INTERNATIONAL TELECOMMUNICATION UNION

TELECOMMUNICATION STANDARDIZATION SECTOR

STUDY PERIOD 2013-2016

STUDY GROUP 13

TD 208 (PLEN/13)

English only

Original: English

Question(s): All/13

TD

Source: Chairman, FG IMT-2020

Title: FG IMT-2020: Report on Standards Gap Analysis

Report on Standards Gap Analysis

Summary

This deliverable provides the report of standards gap analysis as a final output document from ITU-T Focus Group on IMT-2020, FG IMT-2020. Appendices of this deliverable are attached as output documents for five study areas, high-level network architecture, end-to-end QoS framework, emerging network technologies, mobile front haul and back haul, and network softwarization.

Keywords

IMT-2020, Mobile front haul, Mobile back haul, Network Softwarization, QoS, QoE, ICN, CCN, Network Performance, QoS parameters, QoS classes

Introduction

This deliverable is the final output report from Focus Group on IMT-2020, FG IMT-2020, which was established at ITU-T SG13 meeting in April 2015, and worked from June to October 2015. It reports standards gap analysis based on the studies on several key technical topics related non-radio parts of IMT-2020.

Contact: Peter Ashwood-Smith Email: Peter.AshwoodSmith@huawei.com
Huawei Technologies

Table of Contents

1 Scope....4
2 References....4
3 Definitions....4
3.1 Terms defined elsewhere....4
3.2 Terms defined in this Deliverable....7
4 Abbreviations and acronyms....7
5 Conventions....9
6 Executive Summary....9
7 Gap analysis and recommendations to parent group....14
7.1 High-level Architecture....14
7.1.1 Standardization gaps on High-level Architecture....14
7.1.2 Recommendations to parent group on High-level Architecture....19
7.2 Network Softwarization....19
7.2.1 Standardization gaps on Network softwarization....19
7.2.2 Recommendations to parent group on Network softwarization....27
7.3 End-to-end QoS....27
7.3.1 Standardization gaps on end-to-end QoS....27
7.3.2 Recommendations to parent group on End-to-end QoS....30
7.4 Mobile front haul and back haul....30
7.4.1 Standardization gaps on Mobile front haul and back haul....30
7.4.2 Recommendations to parent group on Mobile front haul and back haul....34
7.5 Emerging Network Technologies....35
7.5.1 Standardization gaps on Emerging Network Technologies....35
7.5.2 Recommendations to parent group on Emerging Network Technologies....40
8 Conclusion and future work....40
Bibliography....42
Appendix I: High-level Architecture....**Error! Bookmark not defined.**
Appendix II: Network Softwarization....63
Appendix III: End-to-end QoS....110
Appendix IV: Mobile front haul and back haul....130
Appendix V: Emerging Network Technologies....153

The report of Standards Gap Analysis

1 Scope

This deliverable provides the report of standards gap analysis as a final output document from ITU-T Focus Group on IMT-2020, FG IMT-2020. Appendices[1] of this deliverable are attached as output documents for five study areas: high-level network architecture, an end-to-end quality of service (QoS) framework, emerging network technologies, mobile front haul and back haul, and network softwarization.

2 References

None

3 Definitions

3.1 Terms defined elsewhere

This Deliverable uses the following terms defined elsewhere:

3.1.1 IMT-2020 [b-ITU-R M-2083-0]: systems, system components, and related aspects that support to provide far more enhanced capabilities than those described in Recommendation ITU-R M.1645.

3.1.2 Peak data rate [b-ITU-R M-2083-0]: the maximum achievable data rate under ideal conditions per user/device (in Gbit/s).

3.1.3 User experienced data rate [b-ITU-R M-2083-0]: the achievable data rate (in Mbit/s or Gbit/s) that is available ubiquitously[2] across the coverage area to a mobile user/device.

NOTE - The term “ubiquitous” is related to the considered target coverage area and is not intended to relate to an entire region or country.

3.1.4 Latency [b-ITU-R M-2083-0]: the contribution by the network to the difference in time (in ms). between when the source sends a packet and when the destination receives it.

3.1.5 Mobility [b-ITU-R M-2083-0]: from a performance target point of view, mobility is the maximum speed (in km/h) at which a defined QoS and seamless transfer can be achieved between radio nodes, which may belong to different layers and/or radio access technologies (multi-layer/-RAT).

---

[1] Note: The appendices contain documents that were produced during the FG-IMT 2020 focus group in order to investigate gaps in standardization related to IMT-2020. While the request from SG-13 was to deliver a report outlining standardization gaps, the consensus of the focus group was that the working documents produced and used during the focus group work contained useful information for future work and should be captured. Note, however, the focus group concentrated on producing accurate descriptions of the standardization gaps in the main body of this document; some typographical errors may exist in the appendices. They are, however, the output of the focus group but are provided for information only.

3.1.6 Connection density [b-ITU-R M-2083-0]: the total number of connected and/or accessible devices per unit area (e.g. per km^2).

3.1.7 Energy efficiency [b-ITU-R M-2083-0]: energy efficiency has two aspects:

- on the network side, energy efficiency refers to the quantity of information bits transmitted to and received from users, per unit of energy consumption of the radio access network (RAN) (in bit/Joule);
- on the device side, energy efficiency refers to quantity of information bits per unit of energy consumed (in bit/Joule) by the communication module.

3.1.8 Spectrum efficiency [b-ITU-R M-2083-0]: the average data throughput per unit of the spectrum resource and per cell (measured in bit/s/Hz).

Note: - A cell is a radio coverage area over which a mobile terminal can maintain a connection with one or more units of radio equipment located within that area. For an individual base station, this is the radio coverage area of the base station or of a subsystem (e.g. sector antenna).

3.1.9 Area traffic capacity [b-ITU-R M-2083-0]: the total traffic throughput served per geographic area (in $Mbit/s/m^2$)

3.1.10 future network (FN) [b-ITU-T Y.3001]: A network able to provide services, capabilities, and facilities difficult to provide using existing network technologies. A future network is either:

a) A new component network or an enhanced version of an existing one, or,

b) A heterogeneous collection of new component networks or of new and existing component networks that is operated as a single network.

NOTE – The plural form "Future Networks" (FNs) is used to show that there may be more than one network that fits the definition of a future network.

3.1.11 network virtualization [b-ITU-T Y.3011]: A technology that enables the creation of logically isolated network partitions over shared physical networks so that heterogeneous collection of multiple virtual networks can simultaneously coexist over the shared networks. This includes the aggregation of multiple resources in a provider and appearing as a single resource

3.1.12 software-defined networking [b-ITU-T Y.3030]: A set of techniques that enables to directly program, orchestrate, control and manage network resources, which facilitates the design, delivery and operation of network services in a dynamic and scalable manner.

3.1.13 energy saving within networks [b-ITU-T Y.3021]: This is where network capabilities and their operations are set up in a way that allows the total energy for network equipment to be systematically used in an efficient manner, resulting in reduced energy consumption compared with networks that lack these capabilities and operations.

NOTE – Network equipment includes routers, switches, equipment at the terminating point e.g., optical network units (ONUs), home gateways, and network servers such as load balancers and firewalls. Network equipment is typically composed of various components such as switching fabric, line cards, power supply, and cooling.

3.1.14 cloud service customer [b-ITU-T Y.3501]: A person or organization that consumes delivered cloud services within a contract with a cloud service provider.

3.1.15 cloud service provider [b-ITU-T Y.3501]: An organization that provides and maintains delivered cloud services.

3.1.16 management system [b-ITU-T M.60]: A system with the capability and authority to exercise control over and/or collect management information from another system.

3.1.17 device [b-ITU-T Y.3021]: This is the material element or assembly of such elements intended to perform a required function.

3.1.18 equipment [b-ITU-T Y.3021]: A set of devices assembled together to form a physical entity to perform a specific task.

3.1.19 virtual resource [b-ITU-T Y.3011]: An abstraction of physical or logical resource, which may have different characteristics from the physical or logical resource and whose capability may not be bound to the capability of the physical or logical resource.

3.1.20 logical resource [b-ITU-T Y.3011]: An independently manageable partition of a physical resource, which inherits the same characteristics as the physical resource and whose capability is bound to the capability of the physical resource.

NOTE – "independently" means mutual exclusiveness among multiple partitions at the same level.

3.1.21 resource management [b-ITU-T Y.3520]: The most efficient and effective way to access, control, manage, deploy, schedule and bind resources when they are provided by service providers and requested by customers.

3.1.22 hypervisor [b-ITU-T Y.3510]: A type of system software that allows multiple operating systems to share a single hardware host.

NOTE –Each operating system appears to have the host's processor, memory and other resources, all to itself

3.1.23 virtual machine [b-DMTF OVF]: The complete environment that supports the execution of guest software.

3.1.24 identifier [b-ITU-T Y.2091]: An identifier is a series of digits, characters and symbols or any other form of data used to identify subscriber(s), user(s), network element(s), function(s), network entity(ies) providing services/applications, or other entities (e.g., physical or logical objects).

3.1.25 locator [b-ITU-T Y.2015]: A locator is the network layer topological name for an interface or a set of interfaces. Locators are carried in the IP address fields as packets traverse the network.

NOTE – In Recommendation ITU-T Y.2015, locators are also referred to as location IDs.

3.1.26 node ID [b-ITU-T Y.2015]: A node ID is an identifier used at the transport and higher layers to identify the node as well as the endpoint of a communication session. A node ID is independent of the node location as well as the network to which the node is attached so that the node ID is not required to change even when the node changes its network connectivity by physically moving or simply activating another interface. The node IDs should be used at the transport and higher layers for replacing the conventional use of IP addresses at these layers. A node may have more than one node ID in use.

NOTE – Recommendation ITU-T Y.2015 specifies a node ID structure.

3.1.27 name [b-ITU-T Y.2091]: A name is the identifier of an entity (e.g., subscriber, network element) that may be resolved/translated into address.

3.1.28 domain [b-ETSI NFV MANO]:

Administrative domain is a collection of systems and networks operated by a single organization or administrative authority. Infrastructure domain is an administrative domain that provides virtualized infrastructure resources such as compute, network, and storage or a composition of those resources via a service abstraction to another administrative domain, and is responsible for the management and orchestration of those resources. .

NOTE1: Different networks and different parts of a network may be built as different domains using separate technologies or having different control paradigms,

NOTE2: Different networks and different parts of a network may be owned by a single administration creating an *administrative domain*. Services are enabled and managed *over multiple administrations or over multi-domain single administration*.

NOTE3: A *multitenancy domain* refers to set of physical and /or virtual resources in which a single instance of a software runs on a server and serves multiple tenants. A tenant is a group of users who share a common access with specific privileges to the software instance. A service or an application may be designed to provide every tenant a dedicated share of the instance including its data, configuration, user management, tenant individual functionality and non-functional properties.

3.2 Terms defined in this document

This document defines the following terms:

3.2.1 Network Softwarization: Network softwarization is an overall transformation trend for designing, implementing, deploying, managing and maintaining network equipment and network components by software programming, exploiting characteristics of software such as flexibility and rapidity of design, development and deployment throughout the lifecycle of network equipment and components, for creating conditions that enable the re-design of network and services architectures; allow optimization of costs and processes; and enable self-management.

3.2.2 back haul: the network path connecting the base station site and the network controller or gateway site.

3.2.3 front haul: the intra-base station transport, in which a part of the base station function is moved to the remote antenna site. (Note that this definition is equivalent to the definition given in MEF 22.1.1 for the current 4G technology.)

4 Abbreviations and acronyms

This document uses the following abbreviations and acronyms:

Ed: Not complete.

5G	The fifth generation mobile network
AAA	Authentication, Authorization, Accounting
APN	Application Network
AR	Augmented Reality
BH	back haul
CCN	Content Centric Networking
CGF	Converged Gateway Function
CN	Core Network
CPRI	Common Public Radio Interface
C-RAN	Cloud RAN
CS	Content Store
D2D	Device-to-device
D2N	Device-to-Network
DBA	Dynamic Bandwidth Assignment
DWDM	Dense Wavelength Division Multiplex
E2E	End-to-End

EPC	Evolved Packet Core
FH	front haul
FIB	Forwarding Information Base
GBR	Guaranteed Bit Rate
GW	Gateway
GTP	Generic Tunnelling Protocol
GTP-C	GTP Control
ICN	Information Centric Networking
IDC	Internet Data Center
IMT	International Mobile Telecommunications
IoT	Internet of Things
IP	Internet Protocol
IPDV	IP packet Delay Variation
IPER	IP packet Error Rate
IPLR	IP packet Error Ratio
IPTD	IP packet Transfer Delay
KPI	Key Performance Index
LINP	a logically isolated network partitions
LISP	Location/Identity Separation Protocol
MBR	Maximum guaranteed Bit Rate
MIMO	Multiple-Input and Multiple-Output
MEC	Mobile Edge Computing
MNO	Mobile Network Operator
MPLS	Multi-Protocol Label Switching
MTC	Machine Type Communication
NAS	Non-Access Stratum
NDN	Named Data Networking
NFV	Network Function Virtualization
NP	Network Performance
OAM	Operation, Administration and Management
OBSAI	Open Base Station Architecture Initiative
ODN	Optical Distribution Network
OSU	Optical Subscriber Unit
OTN	Optical Transport Network
PDN	Packet Data Network

PGW Packet Data Network Gateway
PIF Protocol Independent Forwarding
PIT Pending Interest Table
POF Protocol Oblivious Forwarding
PON Passive Optical Network
PTN Packet Transport Network
QCI QoS Class Identifier
QoE Quality of Experience
QoS Quality of Service
RAN Radio Access Network
RAT Radio Access Technologies
RoF Radio over Fiber
TSDN Transport SDN
TWDM Time and Wavelength Division Multiplex
SDN Software Defined Networking
SR Segment Routing
UCF Unified Control Function
UE User Equipment
UHD Ultra High Definition
UMTS Universal Mobile Telecommunication System
UNI User Network Interface
VLAN Virtual LAN (Local Area Network)
VM Virtual Machine
VNF Virtual Network Function
VPN Virtual Private Network
WDM Wavelength Division Multiplex

5 Conventions

Within the context of this document, the term IMT-2020 refers to the technology and networks defined for future mobile networking. Within the telecommunications industry this is commonly referred to “fifth generation mobile networking”, or simply 5G. For the purposes of this document, IMT-2020 and 5G are synonymous.

6 Executive Summary

The purpose of this document is to provide recommends to the ITU-T on the requirements for standardization related to the wireline elements of “fifth generation mobile” (5G) or more properly

referred to as IMT-2020) networks as per the terms of reference of this focus group. It is not a purpose of this document to provide a large amount of tutorial/overview material but instead to reference the copious amounts of such material where necessary and to give sufficient context that experts in the various fields can understand what gaps are being identified.

While the breadth of the document is wide we do not claim to have covered every possible aspect of IMT-2020 wireline networking, indeed IMT-2020 is a large and moving target and therefore we have had to extrapolate in many areas what we believe the wireline requirements will be and what standards are missing or need improvement. There are definitely areas we have missed that will require further study and some of these are outlined as gaps in this gap analysis.

IMT-2020 systems will differentiate themselves from fourth generation (4G) systems not only through further evolution in radio performance but also through greatly increased flexibility end-to-end. This end-to-end flexibility will come in large part from the incorporation of softwarization into every component. Well known techniques such as SDN, NFV and cloud computing will together allow unprecedented flexibility in the IMT-2020 system. Such flexibility will enable many new capabilities including network slicing.

The following figure gives a broad view of key areas that require study to effectively determine all the wireline gaps.

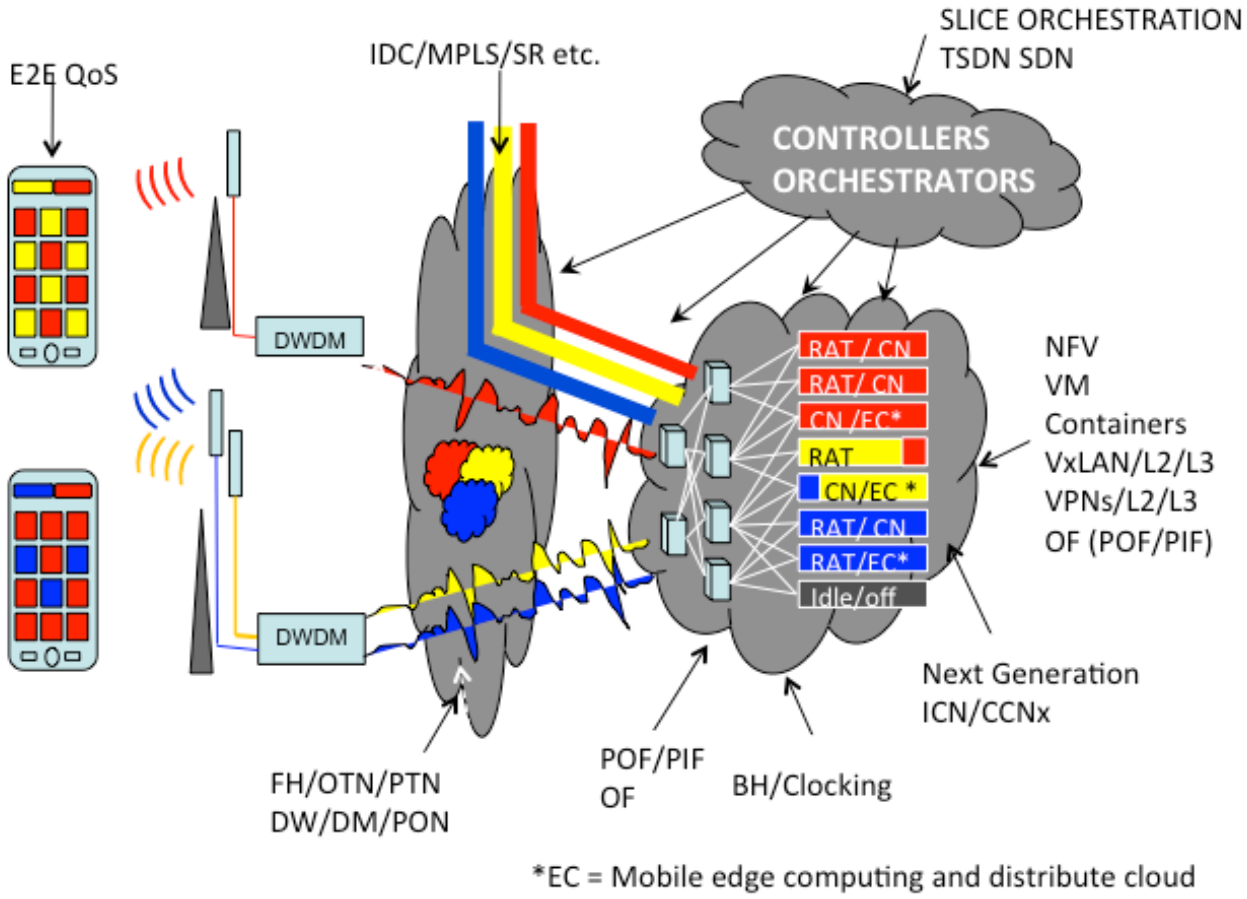


Figure 1: Focus Group IMT 2020 wire line potential gap analysis areas

Since the primary purpose of the focus group is to generate a list of gaps/recommendations the reader will find this list of gaps and recommendations immediately after this executive summary. Following the list of gaps/recommendations the reader will find the detailed work sub topics as appendices.

This focus group has looked at the following wireline aspects of IMT-2020 and has studied each in some detail and produced detailed gaps related to each subject. Due to the short and fixed duration of the Focus Group, there will be some areas, one example is security, which not have been addressed. This should also be considered when formulating possible new work on standardization topics.

A – High level architecture – This study focused on major requirements of IMT-2020 and the analyses of technical gaps to satisfy them. The gaps identified some of the major challenges in IMT-2020 including the diversities in requirements, particularly in bandwidth, mobility, and signalling. The flexibility of the architecture, the tight integration of various radio access networks as well as fixed access networks, end-to-end OAM, among others, are also identified as essential requirements in IMT-2020. Finally, the study provides a high-level network architecture based on the analysis of the requirements, which will be more elaborated or changed in the standardization phases.

Figure 2 provides an overall view of the architecture.

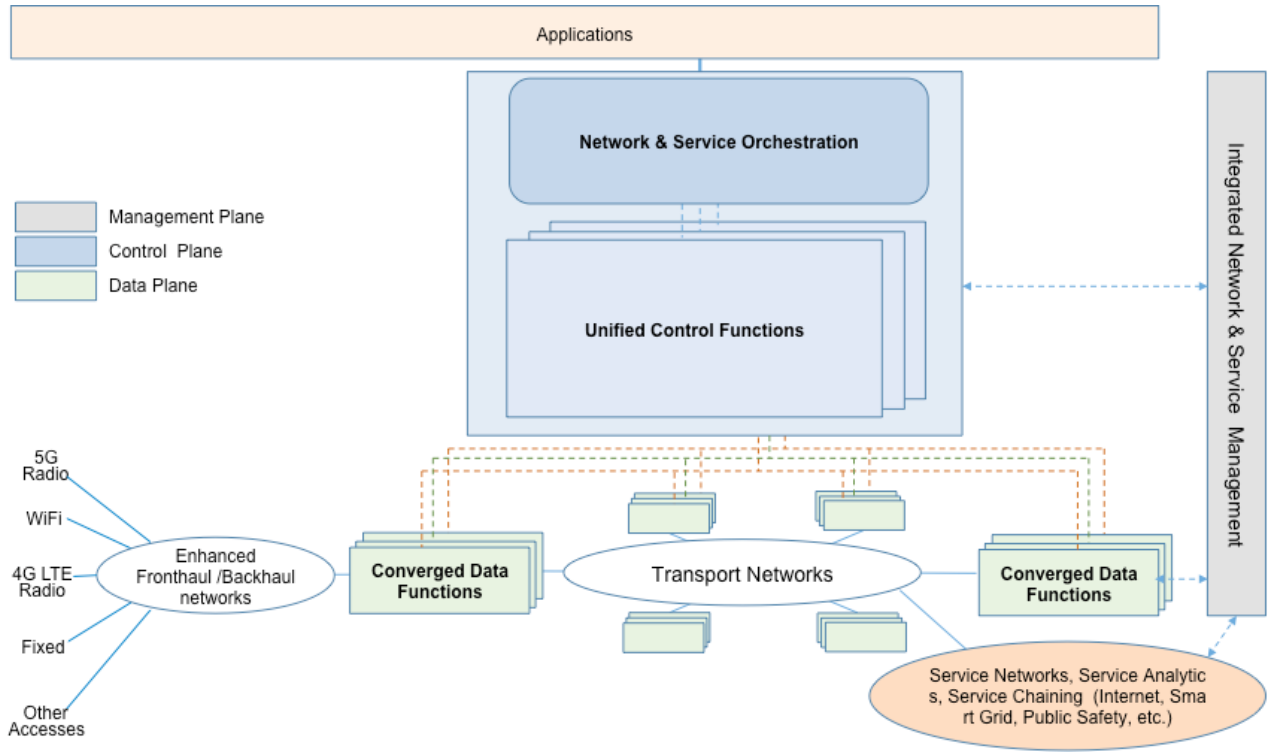


Figure 2: Network architecture for IMT-2020 networks

.

B – Network softwarization – Network softwarization is an overall transformation trend for designing, implementing, deploying, managing and maintaining network equipment and network components by software programming, exploiting characteristics of software such as flexibility and rapidity of design, development and deployment throughout the lifecycle of network equipment and components, for creating conditions that enable the re-design of network and services architectures; allow optimization of costs and processes; and enable self-management. All these bring added value

to network infrastructures. The terminology, Network Softwarization, was first introduced in Academia, at the NetSoft conference in 2015, the first IEEE Conference on Network Softwarization, to include broader interests regarding Software Defined Networking (SDN) and Network Functions Virtualization(NFV),Network Virtualization, Mobile Edge Computing, Cloud and IoT technologies.

Figure 3 provides an overall view of network softwarization in IMT-2020 networks.

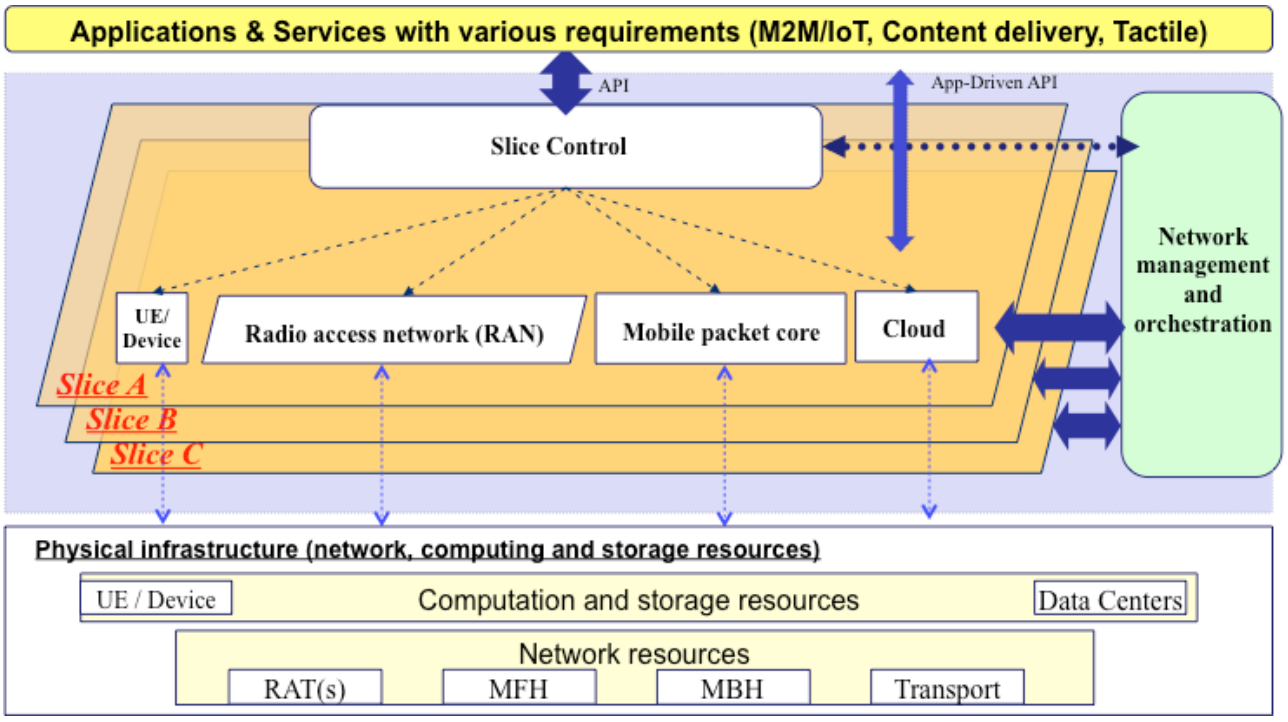


Figure 3: Network softwarization view of IMT-2020 mobile networks

C – End-to-End QoS – This work focused on how the wireline network together with the wireless network can provide guaranteed end-to-end QoS. It provides a survey of various white papers on the subject and identifies differences in how QoS is defined/measured etc. across the different organizations. In addition to a conventional QoS standardization approach, IMT-2020-specific use cases need new approaches in areas of definition of end-to-end connectivity supervision and integrity, QoS parameters, performance objectives, QoS classification, budget allocation, measurement/monitoring methodology, etc. These gaps identify device-to-device/device-to-network QoS requiring additional standardization.

Figure 4 provides an overall view of end-to-end QoS for both wireless and wireline networks.

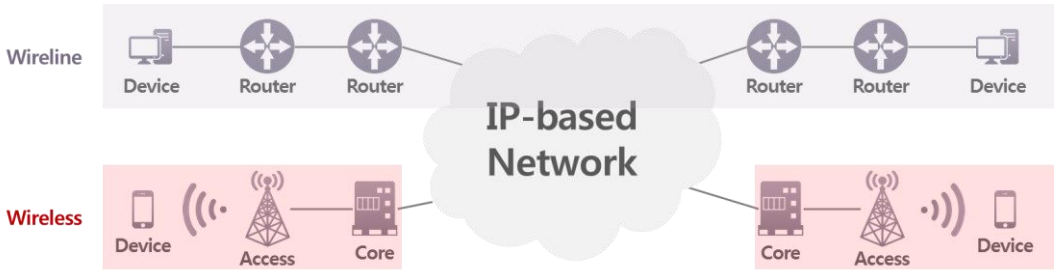


Figure 4: The scope end-to-end QoS standardization for 3GPP (Red) and ITU-T Standards (Purple)

D – front haul/back haul – This work focused on the different transport aspects of the IMT-2020 wireline network. front haul refers to the intra-base-station transport, where part of the base station function is moved to the remote antenna site. Back haul is the packet based communications between the base stations and the various entities that make up the packet processing elements of a core network. This work focuses almost exclusively on the front haul because this is where the majority of the gaps occur. The front haul analysis looks in detail at the projected IMT-2020 bandwidth requirements of a current C-RAN architecture and various architectural variants thereof. Most of the gaps revolve around the need to optimize the front haul network, either to reduce/right-size the bandwidth demands or to increase/optimize the bandwidth supply. An example of the former is to allow front haul to reduce bandwidth capacity when there is not much data being transmitted or received from a given remote site. An example of the latter is to use more advanced transport solutions that are tailored to the front haul application and optimized for low power, low fiber-count, and low cost. The back haul gaps however seem limited to the successful handling of network timing and synchronization, and low latency; as well as improvements in power consumption. All of these suggest gaps in the current standardized technology, many of which are already being addressed by work going on in ITU-T SG15 and various groups in the IEEE. The basic recommendation on FH/FB to SG13 is to establish liaison to all these established groups to better coordinate their efforts.

Figure 5 provides an overall view of the front haul in FMT-2020 networks.

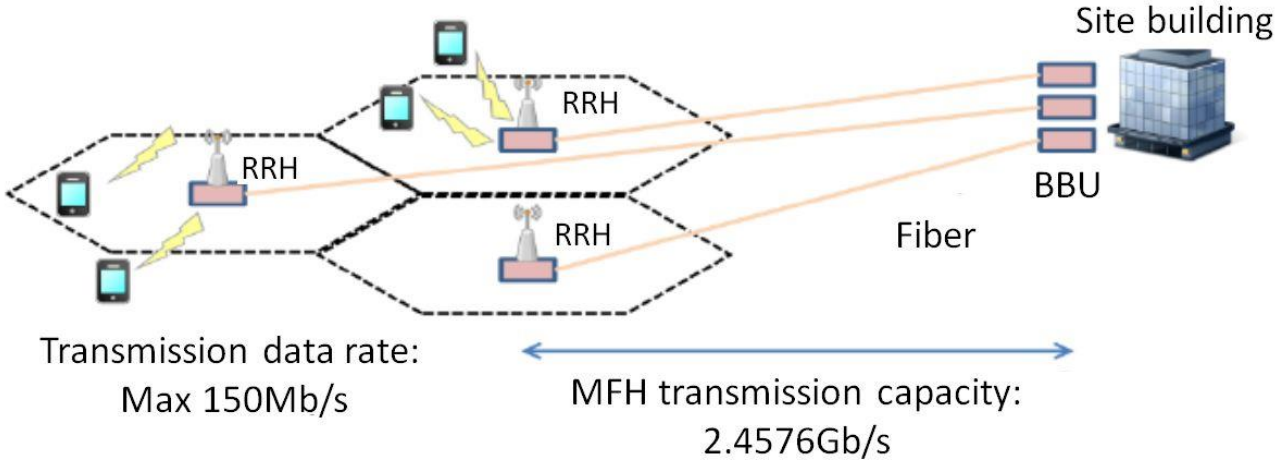


Figure 5: IMT-2020 front haul

E- Emerging Network Technologies – This work focused on the network requirements of IMT-2020 to support enhanced mobile broadband, massive machine-type communications, and ultra-reliable low-latency communications. The work specifically looked at Information Centric Networking (ICN) protocols as possible emerging technologies to meet these needs. Current research shows ICN is possible technology that can provide benefits for IMT-2020 network supporting very large-scale heterogeneous devices, IoT networks, new mobility models, edge computing, and end device self-configuration. Because ICN is a flexible networking architecture, The study scoped the ICN gaps with several incremental deployment options, as described in Appendix V. In clause 7.5, standardization gaps have been identified for ICN deployment in IMT-

2020. The basic recommendation on Emerging Network Technologies such as ICN, to SG13 is to establish liaison to IETF ICNRG, universities and research organizations to further feasibility studies.

7 Gap analysis and recommendations to Study Group 13

7.1 High-level Architecture

7.1.1 Standardization gaps for the IMT-2020 high-level architecture

The gaps included in this clause were extracted from the output of the high-level architecture gap analysis work included in Appendix I.

Gap A.1. Various bandwidth/data-rates demands	Priority: High
Description: The current network architecture is not appropriate to support various bandwidth demands, from ultra-high to extremely low, required for IMT-2020 and beyond. Enhancement of the network architecture should be studied to make IMT-2020 network more flexible and resilient with enhanced capabilities including the upgrades to session or bearer management, more efficient multicast methods, distributed function deployment, network slicing, and efficient codecs.	
Related work: 3GPP, SG13	

Gap A.2. Complex connectivity model	Priority: High
Description: Local offloading can be used to support efficient traffic distribution. However, local offloading requires a separate local mobile GW in existing IMT networks, which also brings up the need to support multiple APN connectivity in user devices or initiate APN switching, which may result in service disruption of on-going sessions and operational complexities. The network architecture for IMT-2020 should be studied to eliminate the need for separate local mobile specific gateways, additional APN and associated signalling.	
Related work: 3GPP, SG13	

Gap A.3. Application-aware and distributed network architecture	Priority: High
Description: The heavily centralized architecture of existing IMT networks is should be changed to cope with the explosion of mobile data traffic. In IMT-2020 networks, therefore, gateways to an IMT-2020 core network can be flexibly located closer to the cell sites, which will bring a significant reduction on back haul and core network traffic by enabling placing content servers closer to mobile devices and also be beneficial to the latency of the services. The IMT-2020 core network, therefore, is envisioned to be a distributed network being composed of the multiple distributed gateways. The architectural changes expected from the distributed network, including the point-to-point access architecture between a UE and a service network in existing IMT network, should be studied.	
Related work: 3GPP, SG13	

Gap A.4. Signalling complexity in massive MTC	Priority: High
Description: The IMT-2020 network should address architectural issues coming from the massive number of MTC devices. The traffic generated by a very large number of connected devices typically will be a relatively low volume of non-delay-sensitive data; however, the traffic is characterized as intermittent short burst traffic. The main problem in supporting the intermittent short burst traffic is that traffic has to go through the full signalling procedure, which causes the waste of battery life, spectrum and network capacity. The enhancement of current monolithic bearer management and the accompanied signalling in IMT-2020 network should be studied to cope with the issues coming from the increase of terminals.	
Related work: 3GPP, SG13	

Gap A.5. Increasing service availability	Priority: High
Note: from second call, confirmation to be confirmed. (via email) Description: IMT-2020 networks should offer several ways to increase service availability by using the ability to replicate content and service functions and the use of forwarding functions for short and long term caching. In addition, delay-tolerant networking aspects, such as cache-and-forward, are very useful in the last mile where the content objects can be opportunistically pushed to or pulled by the end user based on its wireless conditions. The IMT-2020 network architecture should study methods to increase service and content availability.	
Related work: IETF, SG13	

Gap A.6. Signalling to reduce end-to-end complexity	Priority: High
Description: There are various signalling procedures that contribute to the end-to-end connectivity establishment involving all network components such as the radio interface, the front haul/back haul and the mobile core network. Besides the transport delay through the network components, signalling which is basically accompanied in the beginning of each new session or transmission may have more serious impacts on total end-to-end latency. For IMT-2020 and beyond, more efficient signalling protocols or systems should be studied to cope with the limitations on the existing mobile systems.	
Related work:	

Gap A.7. End-to-end network latency model	Priority: High
Description: Latency studies carried out on many IMT-Advanced deployed networks demonstrate that the 3GPP specification provides adequate guidelines, while actual IMT-Advanced network performance varies due to many variables and adjacent ecosystems. Similarly, a network latency	

model and an end-to-end latency budget for services should be studied so that it provides optimal performance for many diverse applications in IMT-2020 networks.
Related work: 3GPP, SG13

Gap A.8. Mobile network optimized softwarization architecture	Priority: High
Description: The softwarization, has been initially designed for wired networks and it may not be optimized for mobile networks. The mobile network optimized softwarization should be studied considering the best use of existing features to overcome the performance issues which may arise in some of SDN solutions.	
Related work: SG13	

Gap A.9. Data plane programmability	Priority: Medium
Description: The requirements of the data plane in an IMT-2020 network will vary depending on the service characteristics of new emerging services such as ICN. In-network data processing and service provisioning are the capabilities to cope with the diverse requirements of the data plane. However, the capabilities have not studied from mobile network perspective. The capabilities to support easier provisioning of new emerging services in IMT-2020 architecture should be studied.	
Related work:SG13	

Gap A.10. End-to-end QoS framework	Priority: High
Description: Discussions on QoS requirements of current IMT networks mostly focus on QoS at the RAN. An end-to-end (i.e. from a user device to another corresponding user device) QoS framework should be considered in the design of IMT-2020 network architecture.	
Related work: SG12, SG13	

Gap A.11. Energy efficiency	Priority: High
Description: Energy efficiency should be considered in the beginning of the design of new IMT-2020 network architecture. While only a subset of total functions in IMT-2020 networks can be virtualized, the function virtualization and its dynamic management of virtualized functions are expected to contribute to the significant saving of energy. Energy efficiency should be defined first and then the measurement method also should be studied for IMT-2020 network. In addition, energy efficiency from life cycle management point of view should be studied.	
Related work: SG5, ISO 14000 Series	

Gap A.12. Enhancement of privacy and security	Priority: High
Description: IMT-2020 networks should have the capabilities of determining and providing the level of privacy and security required of user devices and application services. Therefore, the security and privacy should be well considered in the network architecture design.	
Related work: 3GPP, SG13, SG17	

Gap A.13. Enhancement identity management	Priority: High
Description: As IMT-2020 networks also aim to realize features of future network architectures and enabling services, the goal of identity management and associated security and privacy requirements should go beyond device identity to also include service, content, or other principles. The major functionalities may include identifier (ID) assignment, name translation, ID certification, name resolution, and related key distribution management. This is particularly important when these IDs are administered by third parties which are open to service/control/forwarding functions in the IMT-2020 network, in which case the requirements of the stake-holders have to be taken into account.	
Related work: SG13 Y.2720, SG17, JCA-IdM, ISO/IEC JTC1 SC27	

Gap A.14. Multi-RAT connectivity	Priority: High
Description: Different radio interfaces have been defined by many different standardization organizations, which leads to complex and coarse-level of interworking in existing IMT networks. The signalling on different radio access networks is independent, resulting in inevitable duplications in signalling for attachment, authentication, and mobility in each radio access networks. Although traffic steering and the selection of best access technology in existing IMT networks is considered, more flexible and optimized multi-RAT interworking architecture should be studied, which may also affect the design of IMT-2020 network architecture. The multi-connectivity through the multiple available radio access networks improves the robustness of the network as well as the throughput performance.	
Related work: 3GPP, SG13	

Gap A.15. Fixed mobile convergence	Priority: High
Description: Fixed access networks[3] should be considered as an access network of IMT-2020 to interwork with other radio access networks. A converged access-agnostic core (i.e., where identity, mobility, security, etc. are decoupled from the access technology), which integrates fixed and mobile core, is envisioned as a direction of IMT-2020. Therefore, the IMT-2020 network architecture	

[3] A fixed network means a wireline network. A Wi-Fi network is regarded as another radio access network in this document.

should be studied to support true fixed and mobile convergence ensuring a seamless user experience within the fixed and mobile domains.
Related work: 3GPP, SG13

Gap A.16. Flexible mobility	Priority: High
Description: It is expected that mobility requirements for user devices will vary depending on the device and/or application types. Many user devices are stationary, e.g., smart meters and CPE, even in mobile networks while fast handover is a key feature of most mobile devices and some applications may address the mobility by setting up a new connection automatically with the help of buffering. The signalling procedure in existing IMT networks is heavy and not optimized for some emerging new services such as IoT. An enhanced mobility architecture should be studied to support "context-aware mobility" considering device types, application characteristics, etc. as the context for the mobility.	
Related work: 3GPP, SG13	

Gap A.17. Mobility management for distributed flat network	Priority: High
Description: As the IMT-2020 core network is envisioned to be a flat distributed network, which is composed of the multiple distributed gateways to cope with traffic explosion and latency requirements of applications, mobility management should be studied aligning with those architectural changes.	
Related work: 3GPP, SG13	

Gap A.18. End-to-end network management in a multi-domain environment	Priority: High
Description: Multiple network management protocols in different network domains make it difficult to support unified network operations over multiple network domains. A unified end-to-end network management should be considered to ensure compatibility and flexibility for the operation and management of an IMT-2020 network.	
Related work: SG13	

Gap A.19. OAM protocols	Priority: High
Description: OAM protocols are not standardized in some parts of IMT networks such as the front haul network. Standard OAM protocols should be studied for fault management and performance management between network equipment that may be commonly used across the IMT-2020 network.	
Related work:	

7.1.2 Recommendations to parent group on High-level Architecture

It is recommended that SGs in ITU-T would study and develop specifications of detailed IMT-2020 reference framework, and validate against requirements and solutions for the gaps identified from the viewpoint of IMT-2020 high-level architecture, which are shown in previous sub clause. It is also recommended that SG13 should frequently communicate to the related SDOs, to minimize duplicated work and maximize synergy effects with other SDOs.

7.2 Network Softwarization

7.2.1 Standardization gaps on Network softwarization

The gaps included in this clause were extracted from the output of the network softwarization gap analysis work included in Appendix II.

Note: The gap reference refers to the respective clause in Appendix II.

Gap B.6.2.1: Efficient accommodation of various applications	Priority: High
Description: It is envisioned that such an infrastructure that efficiently supports a diversified set of application requirements across end-to-end paths, ranging from M2M communication, to autonomous and collaborative driving, virtual reality and video streaming, etc. Network softwarization technologies including SDN, NFV and their extensions for supporting IMT-2020 mobile networks are expected to provide slicing capability both in wired and wireless parts of communication infrastructure, so that each slice provides an isolated environment to efficiently accommodate individual applications meeting specific requirements. The slice should be capable of dynamically adjusting resources to meet the application requirements. The network infrastructure is expected to provide extreme flexibility to support those different capabilities with reasonable cost.	
Related work: ITU-T Y.3011, Y.3012, Y.3300, ETSI ISG NFV, Network Functions Virtualization, 3GPP , IEEE SDN	

Gap B.6.2.2-4: Support for emerging network architectures	Priority: Medium
Description: There are both academic and commercial research activities for defining emerging network architectures that do not assume the underlay network runs TCP/IP protocols. A representative example of such network architectures is ICN (Information Centric Networking). Although the current state of ICN could run on top of TCP/IP as an overlay, inherent benefits of the architecture can be achieved if implemented natively, i.e., directly on top of the underlay, e.g., L1 or L2 networks. There exists a gap in supporting such emerging network architecture in the current network technology, especially when it uses such an emerging network architecture in the context of heterogeneous service delivery and function chaining. Network softwarization provides slicing such that such multiple emerging network architectures could be realized within individual slices.	

Related work: ITU-T Y.3011, Y.3012, Y.3300, SG13 Q15

Gap B.6.6.2: Horizontal extension: End-to-end slicing	Priority: High
Description: The scope of the current SDN technology primarily focuses on the portions of the network such as data-centres, mobile and core networks. In IMT-2020 mobile networks, it is necessary to consider end-to-end application quality and enablement through network softwarization platform. Therefore, there exists a gap between the current projection of SDN and NFV technology development and the requirements for end-to-end application quality. The infrastructure for IMT-2020 mobile networks is desired to support end-to-end control and management of slices and the composition of multiple slices, especially with consideration of slicing over RATs and fixed parts of end-to-end paths. This gap has been analyzed against what is defined in [ITU-T Y.3300].	
Related work: ITU-T Y.3011, Y.3012, Y.3300, ETSI ISG NFV, Network Functions Virtualization	

Gap B.6.6.3: Vertical extension: Deep data plane programmability (data plane enhancement)	Priority: High
Description: The current SDN technology primarily focuses on the programmability of the control plane, and only recently the extension of programmability to the data plane is being discussed both in the research community and in ITU-T SG13 but without well-defined use cases. For IMT-2020 mobile networking, there are several use cases for driving invention and introduction of new protocols and architectures especially at the edge of the network. For instance, the need for redundancy elimination and low latency access to contents in content distribution drives ICN at mobile back haul networks. Protocol agnostic forwarding methods such as Protocol Oblivious Forwarding (POF) discuss the extension to SDN addressing forwarding with new protocols. In addition, protocols requiring a large cache storage such as ICN needs new enhancement. A few academic research projects such as P4 [b-P4] and FLARE [b-FLARE] discuss the possibility of deeply programmable data planes that could implement new protocols such as ICN, but there is no standardization activity to cover such new protocols to sufficient extent. Therefore, there exists a gap between the current projection of SDN and NFV technology development and the requirements for deep data plane programmability. The infrastructure for IMT-2020 mobile networks is desired to support deeper data plane programmability for defining new protocols and mechanisms. This gap has been analyzed against what is defined in [ITU-T Y.3300].	
Related work: ITU-T SG13 Y.3011, Y.3012, Y.3300	

Gap B.6.6.4: Considerations for applicability of softwarization	Priority: High

Desrciption: SDN and NFV are primarily motivated by OPEX and CAPEX reduction and flexible and logically centralized control of network operations, and these technologies aim to focus on softwarization of everything everywhere possible to meet various network management and service objectives. Also the traffic classification is often per flow basis.

In IMT-2020 mobile networks, some applications have stringent performance requirements such as ultra-low latency and high peak rate while others may not require cost-effective solutions. Solutions exist ranging from application driven software-based solutions executed on a virtualization platform with a hypervisor, container or bare metals, to complete hardware-assisted solutions. The former may need performance enhancement enabled by hardware-assisted solutions, while the latter may be facilitated by software-based solutions.

The infrastructure for IMT-2020 mobile network must support traffic classification performed not only by flow-basis but also by other metrics and bundles such as per-device and per-application basis so that software /hardware based solutions may be applied appropriately for individual use cases. Therefore, there exists a gap between the current projection of SDN and NFV technology development and the requirements for applicability of softwarization. This gap has been analyzed against what is defined in [ITU-T Y.3300].

Related work: ITU-T Y.3011, Y.3012, Y.3300, ETSI ISG NFV, Network Functions Virtualization

Gap B.6.6.5: End-to-end reference model for scalable operation	Priority: Medium

Description: Intensive studies are required on both the dimension and the dynamic behaviour of softwarized networks, since such highly virtualized systems will have an enormous number of instances and reactions are not easy to extrapolate from current physical systems.

The virtualized resource handling must be the essential part of the scalable and novel operation architecture, which potentially improves conventional network operations and, possibly even up to the level of supporting disaster recovery, by using softwarized network resiliency and recovery of /with the virtualized systems both in a single domain and in multiple domains.

One of the benefits of IMT-2020 systems should be the end-to-end QoE management, however, this capability will be established on the complex interaction between the virtualized systems including UEs, Cloud Systems, Applications, and the softwarized network systems. The softwarized network system itself will be composed of various virtualized subsystems. An appropriate end-to-end reference model and architecture should be intensively investigated for such complex systems.

.

Related work: None

Gap B.6.6.6: Coordinated APIs	Priority: Medium

Description: In IMT-2020 mobile networks, it may be useful to define APIs so that applications and services can program network functions directly bypassing control and management to optimize the performance, e.g., to achieve ultra-low latency applications. Information modelling should be the most significant issues for the APIs development. It should include virtual resource characteristics, relationship between various resources, operational models, and so on.

Discussions on the programmable interface capabilities should include:

- A level of abstraction sufficient both for system operations and for customization of the capability provided by the interfaces
- Modelling for the virtual/abstracted resource in a multiple-technology environment
- Ease of programming for service and operation velocity
- Technologies for automatic and/or autonomic operations
- Provisioning of classified functional elements suitable for a range of system developers such as supplication service providers, network service providers, and network management operator

These issues should be considered as a gap to be discussed for possible standardization items.

Related work: None

Gap B.6.7: Energy management aspects of network softwarization	Priority: High

Description: Energy-conscious IMT-2020 single domain: optimizing the energy consumption within the limits of a single domain, based on system virtualization and the optimal distribution of VMs as well as M2M scenario This will be coupled with the dynamic adaptation of active and stand-by servers/network functions and the load optimization per active server. A new monitoring framework to measure the energy consumption per server module/networking component and activate low-power states on devices would be needed.

A Group of energy-conscious IMT-2020 domains: optimizing the cumulative energy consumption in a group of domains, based on optimal distribution of VMs across all of the servers that belong in the group of domains using policy-based methods. Measuring the energy consumption on the domain level and deploying policies and solutions that will achieve decreased cumulative power consumption across the whole group of domains would be needed.

Related work: None

Gap B.6.8: Economic incentives aspects of network softwarization	Priority:

Description: Sufficient attention needs to be paid to economic and social aspects such as economic incentives in designing and implementing the IMT-2020 network softwarization and its architecture in order to provide a sustainable competition environment to the various participants.

Drastic reduction of the operational and lifecycle costs for all components and systems involved in the IMT-2020 network softwarization would be recommended for efficient and sustainable deployment enabling an appropriate return for all involved actors.

Ways of resolving economic conflicts including tussles in the IMT-2020 networking and servicing ecosystem that include an economic reward for each participant’s contribution are

becoming essential. Different participants may pursue conflicting interests, which could lead to conflict over the overall multi-domain operation of network softwarization and controversy in international/domestic regulation issues.
Related work: Y.3013

Gap B.7: Network management and orchestration	Priority: High
Description: There are two aspects to consider for the network management and orchestration for the network softwarization. The first aspect is how to manage and orchestrate the softwarized network components. The second is how to softwarize network management and orchestration functionality. The current technology gaps to be filled in are provided.	
Related work: ITU-T Y.SDN-ARCH, Y.AMC, ETSI ISG NFV, MANO, TMF ZOOM	

Gap B.8.1: Support enhanced MEC management	Priority: Medium
Description: Mobile edge computing (MEC) uses a virtualization platform for applications running at the mobile network edge. Although mobile edge server lifecycle management is supported by existing NFV management functionality, MEC management should support some enhancements in following aspects: 1) Mobile edge application lifecycle management: The MEC management functionality should support the instantiation and termination of an application on a Mobile edge server within the Mobile edge system when required by the operator or in response to a request by an authorised third-party. 2) Mobile edge application service management: The Mobile edge platform provides services that can be consumed by authorised applications. Applications should be authenticated and authorised to access the services. The services announce their availability when they are ready to use, and mobile edge applications can discover the available services.	
Related work: ETSI MEC ISG GS MEC 002 &GS MEC 003, ETSI NFV ISG IFA WG	

Gap B.8.2: Support inter-edge mobility of a MEC system	Priority: High
Description: Mobility is an essential component of mobile networks. Considering some mobile edge applications are specifically related to the user activity, the MEC system needs to maintain some application-specific user-related information that needs to be provided to the instance of that application running on another Mobile edge server. Therefore a MEC system should support an inter-edge mobility mechanism for service continuity when the user is moving to an area served by another mobile edge platform that hosts the application.	
Related work: 3GPP TS 23.401, TR 23.714, ETSI MEC ISG GS MEC 002&GS MEC 003	

Gap B.8.3: Support more simple and controllable APIs of a MEC system	Priority: Medium
Description: In order to enable the development of a strong ecosystem for mobile edge computing (MEC), it is important to develop APIs that are as simple as possible and directly meeting the needs of applications. In addition, standardized APIs may need some enhancement to provide radio analytics or radio network information. Therefore, the MEC system should optimize existing APIs to make them more simple and controllable.	
Related work: ETSI ISG MEC GS MEC 003, 3GPP RAN	

Gap B.8.4: Support traffic routing among multiple MEC applications	Priority: High
Description: The mobile edge platform routes selected uplink and/or downlink user plane traffic between the network and authorized applications, and between authorized applications. More than one applications might be selected for the user plane traffic to route through properly (e.g. video optimization and Augumented reality). The MEC system should support a traffic routing mechanism among multiple applications: selection and routing during traffic redirection based on re-direction rules which are defined by the operator per application flow, and selected authorized applications can modify and shape user plane traffic.	
Related work: 3GPP TR 23.718, TS 23.203, ETSI MEC ISG GS MEC 002, GS MEC 003	

Gap B.9: Distributed cloud for service provider	Priority: Medium
Description: The existing IMT network has its limitation and lacks of flexibility and agility for deploying the network functions and applications at any location where the performance and user experience would be optimized. In order to meet the extremely various demands of the services in IMT-2020, for example, from ultra-low latency to high-latency tolerable service, distributed cloud technology provides a viable solution. To realize its benefits, the following gaps would need to be filled: (1) Distributed storage services that provide uniform, system-wide, distributed block storage for the applications (e.g., OpenStack's Swift and Cinder subsystems) (2) Networking services that SDN enables such as cloud-wide, virtualized connectivity, both at L2 and L3 levels, such as OpenStack's Quantum, (3) Distributed compute services that manage VMs, doings tasks such as start, stop, migration and supervisions of VMs is to be performed by a cloud computing service (e.g., CloudStack and OpenStack's Nova) and (4) Cloud management API for applications on top of the cloud infrastructure (e.g., OpenStack) for application deployment, migration and portability. Although it is ideal to have all applications running inside VM's, reality, at least in the short term, dictates that some tasks must continue to execute on non-virtualized or specialized hardware. In order to limit the extra OPEX burden such system anomalies represent, it is still necessary to provide these environments with an API that makes it possible to manage them the same way (i.e.	

by abstracting and presenting a uniform interface to the applications) as is done with VMs, so that it is still possible to keep parts of the management APIs (loading, start, stop etc.) uniform and identical to the ones as in VMs.
Related work: ETSI ISG NFV, OpenStack, OpenDayLight, CloudStack

Gap B.10: In-network data processing	Priority: Medium
Description: One use case scenario of in-network data processing is included in ITU-T SG13 that deals with requirements and architecture with in-network data processing. However, only a limited number of use case scenarios are described for network data processing. Further discussion for viable in-network data processing for IMT-2020 mobile network is necessary.	
Related work: SG13/Q15	

Gap B.11: Resource usage optimization	Priority: Medium
A large portion of digital data is transferred repeatedly across networks and duplicated in storage systems, which costs excessive bandwidth, storage, energy, and operations. Thus, great effort has been made in both areas of mobile and fixed networks and storage systems to mitigate the redundancies. However, due to the lack of the coordination capabilities, expensive procedures of C-H-I (Chunking, Hashing, and Indexing) are incurring recursively on the end-to-end path of data processing. Redundancy reduction methodology in an end-to-end path of a softwarized network may be needed for resource usage optimization.	
Related work: IETF CDNI WG, IRTF/ICN-RG	

Gap B.12: Resource abstraction	Priority: Medium
Description: This is no common model that can provide abstraction of various capabilities supported by physical resources that constitute end-to-end scope and are not covered by existing networks, including, physical radio interfaces, packet forwarding and routing in access networks. The granularity of the current abstraction model may not be sufficient to support various approaches to satisfy end-to-end quality requirements of the application, while minimizing impact on utilization of networks.	
Related work: ITU-T Y.3300, ETSI ISG NFV There exists the general guideline for resource abstraction described in [ITU-T Y.3300], specifically, Common resource abstraction model and Granularity of abstraction.	

Gap B.13: Migration towards newly emerging network	Priority: Medium

Description: Network virtualization, described in [ITU-T Y.3011], allows the network providers to integrate legacy support and keep backward compatibility by allocating existing networks to LINPs (i.e., slices) for deploying new network technologies and services or migrating to new network architecture. It is expected that network softwarization, especially slices, can provide migration paths to newly emerging network architectures since it may be possible to accommodate multiple network architectures in slices concurrently. However, there is yet no activity observed for discussing the detailed migration scenario.
Related work: ITU-T Y.3011

Gap B.14: RAN virtualization and slicing under software control	Priority: High
Description: Virtualization of the RAN domain in conjunction with software control is expected to be an effective solution to provide appropriate QoE for diversified service requirements in dynamic way with a reasonable cost. RAN resources and functionalities are mapped onto the network slices in association with service profiles. Following elements should be defined. (1)Slice management and the arrangement of VNFs, virtual topology and software based transport control on the slice. (2)Activation of slice attributes such as the application drive, resiliency, OAM, and security on each slice and inter-slice. (3)Appropriate APIs in some network elements such as UE, X-haul, TSDN, NVFs, OTN/DWDM.	
Related work: ITU-T/SG13,SG15, 3GPP/SA2,SA5 (ITU-T Y.3300, ITU-T Y.3320, ITU-R REP M.2320, 3GPP TR 22.891, 3GPP TS 28.500, 3GPP TR 32.842), in order to introduce the Slice concept and software elements controlled around RAN in E2E including front haul/back haul	

Gap B.15: Capability Exposure	Priority: High
Description: An NGMN 5G White paper has proposed the requirements on network capability exposure to enable business agility, and envisioned a IMT-2020 architecture that includes network capability exposure as a key functionality. 4G network architecture enhancement for capability exposure is an ongoing study in 3GPP SA2. IMT-2020 capability exposure work is on the stage of service requirement research in 3GPP SA1. It mentions that the network slicing capability in the IMT-2020 era could be exposed to the customer by providing to it the specific network slice according to its demand. However, the detailed discussion on capability exposure is not yet done and current use cases are not comprehensive and systematic. The existing ITU-T specifications have covered some aspects in NGN context [ITU-T Y.2234 and ITU-T Y.2240 from a capability perspective], The following points should be studied: • Scenarios and requirements of network capability exposure • Architecture, mechanism and API of capability exposure – Overall architecture for the capability exposure – Potential solutions and the E2E procedure to enable each capability to fulfill the specific service demand – Open APIs interworking with third parties based on the investigation on API work of this document.	

– Privacy and Security
Related work: NGMN 5G White paper has proposed the requirements on network capability exposure. 4G network architecture enhancement for capability exposure is an ongoing study in 3GPP SA2. 3GPP SA1, ITU-T Y.2234 and Y.2240 from a capability perspective.

7.2.2 Recommendations to parent group on network softwarization

(1) Each identified gap needs careful discussion and consideration as to which SDO should pick up the work and whether standardization immediately required or if further discussion towards standardization is needed.
(2) If each identified gap is to be standardized, it needs coordination among related SDOs such as ETSI ISG NFV, ETSI ISG MEC, 3GPP, and ONF.
(3) It is recommended that the network softwarization study should be co-developed further in accordance with other study issues from the following perspectives:

ICN group: deep data plane programmability for implementation and slicing for accommodation

front haul/back haul group: RAN virtualization

High-level Architecture group: overall design of architecture

E2E QoS group: End-to-end application quality

7.3 End-to-end QoS

7.3.1 Standardization gaps on end-to-end QoS

The gaps included in this clause were extracted from the output of the end-to-end QoS gap analysis work included in Appendix III. Note, the gap reference refers to the respective clause(s) in Appendix III.

Gap C.7.3-1. Definition of end-to-end	Priority: High
Description: 3GPP's concept of "end-to-end" comprehensively covers the whole network from a user's device to another user's device. However, its UMTS bearer concept is limited to an interval starting from user's device to PDN gateway (a gateway in wireless core network) for the sake of practicality (i.e., a network operator can influence only its network and its radio interface). (3GPP TS 23.107, TS 23.401 Rel.12)	

ITU-T, on the other hand, attempts to identify network QoS from end-user to end-user by defining UNI to UNI objectives in Y.1541. However, the concept is usually applied to wireline IP-based services without any specific discretion on technologies of lower layers. IMT-2020 QoS standard should define a common single end-to-end definition.
Related work: ITU-T Y.1540, Y.1541, Y.1542, 3GPP TS23.107, TS23.401

Gap C.7.3-2. End-to-end connectivity for D2D/D2N – integrity and supervision	Priority: High
Description: Existing standards for mobile (e.g. 3GPP) and fixed (e.g., ITU-T) networks have been developed for human-to-human and human-to-machine connectivity which is concatenated through the device, access network, core network and server and vice versa. IMT-2020 QoS standard should study device-to-device and device-to-(edge) network connectivity cases, which are generally shorter than conventional connectivity	
Related work: ITU-T I.350, I.356, Y.1540, Y.1561, Y.1563	

Gap C.7.3-3. Different QoS classification among mobile and fixed networks	Priority: Medium
Description: While mobile network-related standards (e.g. 3GPP) specify 13 QoS Classification Indicators (QCI), fixed network-related standards (e.g., ITU-T) introduce 6 QoS classes with different parameters and performance objectives. IMT-2020 QoS standards should study the way to be applicable for both networks.	
Related work: ITU-T I.356, Y.1541, 3GPP TS 23.107	

Gap C.7.3-4. Additional QoS parameters	Priority: High
Description: Latency is just one of parameters to define QoS aspects. An IMT-2020 QoS standard should study other parameters, such loss ratio, delay variation (jitter), etc. for the delivery performance viewpoint,.New parameters should be considered to support IMT-2020 specific use cases for service execution capability such as remote surgical operation, autonomous driving and virtual reality. Also, the impact of new network architectural aspects should be taken into account; network softwarization (e.g. slicing), ICN etc.	
Related work: ITU-T I.350, I.356, Y.1540, G.1010, 3GPP TS23.107	

Gap C.7.3-5. Measurement and monitoring	Priority: Medium
Description: For Device-to-Device and Device-to-Network connectivity cases with very low delay (e.g. 1ms) require definition of the methodology of measurement, reference points and monitoring methodology. An approach using OAM technology for intrusive measurements should be also taken into account for this purpose. While Gap C.7.3-2 focuses on the definition itself, this gap is related to how to manage and operate Gap C.7.3-2.	

Related work: ITU-T I.356, O-series, Y.1541

Gap C.7.3-6 QoS budget allocation for mobile and fixed networks	Priority: Medium
Description: Performance objectives in existing standards (ITU-T & 3GPP) were developed focusing on its own network's connectivity (i.e. mobile or fixed). End-to-end performance objectives covering mobile and fixed networks should be allocated into media-dependent way such as fiber optics and radio etc. Device-to-network communication is different from conventional human-to-human communication in aspects such as frequency of communication (periodic) and type of traffic generated (usually more signalling traffic than data). Device-to-device communication also is distinctly different from the conventional communication because the distance will be much shorter and the configuration will be simpler (with smaller number of nodes). In-depth study is necessary to develop QoS budget allocation for these connectivity configurations.	
Related work: ITU-T Y.1541, 3GPP TS23.107, TS23.401	

Gap C.7.3-7. Performance objectives	Priority: High
Description: The conversational voice application has been considered to have the most stringent performance objectives; end-to-end one way latency of 150ms for human's mouth-to-ear connectivity and 100ms for UNI-to-UNI. Assuming that the revised end-to-end connectivity for device-to-device and device-to-network impose stringent performance objectives, new QoS performance objectives may be required.	
Related work: ITU-T G.1010, ITU-T Y.1541	

Gap C.7.3-8. Layered approach	Priority: Low
Description: Realizing QoS requirements must be based on the structure of technologies and protocols in different layers, ITU-T's Y.1540 standard provides a layered model of performance of IP service to illustrate the point aforementioned. The lower layers do not have end-to-end significance (i.e., it transfers packet from one point to another) but the type of technology employed (e.g., Ethernet-based leased lines) may affect the performance. 3GPP's bearer acknowledges the effect of various layers on IP services, but defines the bearer on layer 1 and 2 for the use of higher layers (3GPP TS 23.107 & 23.401). Nevertheless, both 3GPP and ITU-T acknowledge that the frame work must take into account the impact from performance of layer 1 and 2 in both wireline and wireless media. Higher layers implemented in service execution systems (security, mobility, interworking etc) may also affect performance. The IMT-2020 QoS standard development should study the overall layered structure and inter-relationship.	
Related work: ITU-T Y.1540, 3GPP TS 23.107, TS 23.401	

Gap C.7.3-9. Overall QoS study applicable to IMT-2020	Priority: High
Description: Since new technologies are required to implement the IMT-2020 network, the operational aspects at the QoS level in the real field, and an understanding of QoS end-to-end require the initiation of study of these new concepts (for example, network softwarization (e.g.slicing) and other areas (including, for example network management/OAM, signalling, network architecture, implementation scenarios, etc.) Moreover, the hybrid mobile and fixed network environment of IMT-2020 calls for a systematic and integrated approach to establish a common framework for QoS standards.	
Related work: SG2, SG11, SG12, SG13, SG20 – related recommendations	

7.3.2 Recommendations to parent group on End-to-end QoS

Gaps from C.7.3-1 to C.7.3-8 could be delivered to Study Group 12 for further in-depth standardization. However, for the purpose of operation and management, development of the overall QoS end-to-end standards from the network point of view should be kept inside Study Group 13.

7.4 Mobile front haul and back haul

7.4.1 Standardization gaps on Mobile front haul and back haul

The gaps included in this clause were extracted from the output of the mobile front haul and back haul gap analysis work included in Appendix IV. Note, the gap reference refers to the respective clause(s) in Appendix IV.

Gap D.7.1-1. Large capacity transmission	Priority: Low
Description: There is requirement for large capacity transmission to support ultra high speed mobile transport. Data compression for radio interfaces could be a candidate.	
Related work: CPRI, OBSAI and others have specified 9830.4Mbit/sec for 20MHz, 2x2MIMO x 4ch .and are considering to producing a future specification of 12Gbit/sec. It is recommended that they continue this work toward further enhancement of the specification.	

Gap D.7.3-1. Timing requirements	Priority: Medium
Description: More precise definition and specification of timing requirements is needed, including a breakdown of the impairment budget over the envisioned networks, both for front haul and back haul.	

Related work: Q13/15 is active on these topics, having produced G.827x G.826x series, but not enough for IMT-2020 timing requirements. It is recommended they continue this work.

Gap D.7.3-2. Low latency	Priority: High
Description: Specification of E2E delay of the routing path, processing time and the time in the queues are needed. Related studies, such as, routing of end-to-end communication, processing on the BS instead of the processing on the servers, and minimizing the time in the queues are also expected.	
Related work: For routing of E2E communication, and for minimizing the time in the queue in PON systems, VLAN and DBA protocols are specified in G.987.1 or G.989.1. There are no specifications about interworking with RAN or protocols for non-PON systems.	

Gap D.7.4. Power saving by sleep or rate control	Priority: Medium
Description: For the power saving, sleep control or transmission rate control is expected for the both of radio-over-fibre(RoF)/ Non-radio-over-fibre mobile traffic transmission. Especially, when RoF is used in the front haul, signal is on all the time, even if the data traffic is zero in the ODN.	
Related work: SG15 has delivered G.suppl 45 (mobile use isn't expected) including the power saving for PONs. However sleep or rate control does not seem to be studied which are obtained by the information from other functions.	

Gap D.7.5-1. PON as the virtual digital wireline service	Priority: High
Description: A protocol is needed for CPRI over Ethernet and/or redefinition of function splitting. A study of different function splitting points is needed to optimize the front haul wireless network.	
Related work: Discussion has been started in academia, and also in the IEEE NGFI project, but it has not been standardized yet. Collaboration is expected between wireless and fixed network SDOs.	

Gap D.7.5-2. Large number of fibers for front haul	Priority: High
Description: The front haul with large number of small cells requires a large number of fibers. The PON/ODN/WDM can reduce the number of fibers and interfaces in the base station in the FH.	
Related work: SG15 has delivered G.989.1 which is considering the PON/ODN/WDM systems for BH. Q2/15 is now studying PON/ODN for front haul with RoF. However the use case for front haul with non-RoF or CPRI does not seems to be considered. Q6/15 is now studying G.metro, with looks at similar issues in a metro transport network context.	

Gap D.7.6. Reliability and resiliency	Priority: Medium

Description: Automatic resource reallocation mechanisms and interfaces to negotiate the allocation between the mobile systems, network controller and FH transport systems are expected. Resource reallocation for softwarized front haul system is also expected.
Related work: For PON systems, switching time for wavelengths is specified in G.989.2, and duplex architecture is specified in G.983.1. There are no specifications about automatic resource reallocation mechanisms and interfaces to negotiate the allocation. For optical transport networks (OTN), protection is described in G.873.1 and G.873.2. For packet transport, protection is considered in G.8031 and G.8032. All of these could be leveraged.

Gap D.7.7-1. Diversified types of terminals	Priority: High
Description: Interfaces, required for various terminals and various devices to negotiate with the gateway of IoT devices, service servers or network controllers, are expected.	
Related work: For providing of virtualized networks to the various types of terminals or services, VLAN and DBA protocols are specified in G.987.1 or G.989.1. There are no specifications about interworking with the RAN.	

Gap D.7.7-2. Diversified types of traffic	Priority: High
Description: Interfaces between the mobile systems and front haul/back haul systems are expected in order to monitor the traffic type, the required service policies and the network controller.	
Related work: For providing of virtualized network to isolate traffic based on service, priority or some other policies, VLAN and DBA protocols are specified in G.987.1 or G.989.1. There are no specifications about interworking with RAN.	

Gap D.7.7-3. Diversified types of network operator	Priority: High
Description: Interfaces between mobile systems (or network controllers) and front haul/back haul systems need to allow the operator access to part of the management system.	
Related work: For providing of virtualized network to isolate traffic by the operator, VLAN and DBA protocols are specified in G.987.1 or G.989.1.	

Gap D.7.7-4. Diversified types of RAN	Priority: High
Description: Interfaces between mobile systems (or network controllers) and front haul/back haul systems need to allow access to part of the management system for management of multiple, different RANs.	
Related work: For providing of virtualized network with 3G/4G/IMT-2020/WiFi heterogeneous seamless integrated operation, VLAN and DBA protocols are specified in G.987.1 or G.989.1.	

Gap D.8.1-1. Optimization of module or chip device design	Priority: Medium

Description: For cost reduction of high-speed CPRI (40G/100G) interface, adoption of commodity modules or devices like Ethernet-PHY's is expected. Therefore the standard of functionality or interfaces for CPRI over Ethernet -PHY is needed.
Related work: IEEE1904.3 is starting to discuss on the topic of Radio over Ethernet but has not been standardized yet".

Gap D.8.1-2. PON with WDM overlay	Priority: Low
Description: WDM multiplexing over fiber access systems may be useful for front haul, and these systems need to be described in terms of their system architecture, link parameters, and management features.	
Related work: Q2/SG15 has the G.989 series, that describes a WDM overlay PON system in detail. Q6/15 has the G.metro project, that describes a similar system for the metro transport space.	

Gap D.8.2. Analog radio over optical fiber transmission	Priority: Low
Description: Analog radio-over-fibre (RoF) system requirements, transceiver specifications, and transmission convergence schemes need development, with the goal of producing interoperable analog transport.	
Related work: Q2/SG15 has produced G.sup.55 that describes RoF generically, and has started a project to develop a G.RoF recommendation series. WDM transport systems have been described in Q2/15 and Q6/15.	

Gap D.8.3. CPRI over OTN	Priority: Medium
Description: The transport of CPRI or CPRI-like signals over OTN has demanding requirements, mainly regarding symmetry and the transport of timing for these client signals.	
Related work: Q11/15 has produced G.sup.56 that describes the transport of CPRI over OTN. More work to resolve the timing issues is needed.	

Gap D.8.4. Improved CPRI	Priority: Medium
Description: The existing CPRI format is restrictively simple, and the interface is exacting in its specifications. New formats that solve these problems would be a useful addition.	
Related work: Q2/15 has defined a transcoding method in G.989.3 to reduce the 20% overhead of 8b10b code. More work by the groups that defined CPRI is needed.	

Gap D.8.5. Radio over packet	Priority: Medium
Description: The transmission of radio data over packet networks needs to be studied to determine if it is feasible, and if so, to describe the system as a whole, the network node behaviours and the end-point adaptation functions.	
Related work: IEEE 802.1CM will develop a profile for front haul over Ethernet bridges; IEEE TSN is working on some solutions for precision timing and low-delay packet forwarding over	

Ethernet bridge. IEEE1904.3 is starting to discuss on the topic of radio over Ethernet. Also, the IEEE Comsoc NGFI group is starting work on this and other areas.

Gap D.8.6. Developing function splitting of front haul network	Priority: High
Description: A study of different function splitting points is needed to optimize the front haul wireless network. The key trade-off is the centralization of processing / wireless performance vs. transport bandwidth requirement.	
Related work: Discussion has been started in academia, and also in the IEEE NGFI project,but it has not been standardized yet". Collaboration is expected between wireless and fixed network SDOs.	

Gap D.8.7. Extension of G.metro for the transport of CPRI in MFH/MBH networks	Priority: High
Description: The transport of CPRI or CPRI-like signals in metro networks is currently discussed in Q6/15 where a new draft Recommendation G.metro is being developed. G.metro is a WDM based transport network, where low cost wide tuneable lasers are indispensable in the development of such a recommendation. Photonic integration could also be beneficial. G.metro could be extended to transport CPRI over the MFH/MBH networks	
Related work: Q6/15 has been working on G.metro since March/April 2014, and is targetted for consent in Sep 2016.	

Gap D.9.2.1. Coordination of power saving across MFH/MBH/Radio System	Priority: Medium
Description: For power saving, interworking is expected between the FH, BH and radio systems. Generally flow routes or radio systems (IMT-2020, WiFi, or stations) to the UE are selectable.	
Related work: SG15 has delivered G.9802 with section 9.3 "wavelength resource administration", however interoperability is not addressed.	

Gap.D.9.2.2. Power saving by resource optimization	Priority: Medium
Description: For the power saving, resource optimization is expected such as the optimization of the number of activated OSUs in TWDM PON, the number of allocated optical paths or allocation of wireless resource.	
Related work: SG15 has delivered G.989.1 with section 9.6 "power reduction" for NG PON2. However schemes and interfaces for the FH, BH and radio system do not seem to be studied for the optimal resource allocation.	

7.4.2 Recommendations to parent group on Mobile front haul and back haul

As the result of gap analysis and a survey of the related technologies and standards, 22 standards gaps were identified for front haul/back haul for IMT-2020. These gaps can be categorized into mainly three groups

(1) The gaps depending on the transport technologies which are better to be discussed by the specialists for the transport layer like the SG15:
D.7.3-1, D.7.3-2, D.7.5-2, D.7.6, D.8.1-2, D.8.2, D.8.3, D.8.7.
(2) The gaps related to the network architecture, interworking technologies between the front haul/back haul/Core/Radio-Systems and network softwarization technologies which should be discussed by the collaborations of the transport and architecture groups:
D.7.7-1, D.7.7-2, D.7.7-3, D.7.7-4, D.9.3
(3) The gaps other than previous 2 categories which should be studied in collaboration with several groups such as ITU-R, 3GPP and IEEE:
D.7.1-1, D.7.4, D.7.5-1, D.7.7-4, D.8.1-1, D.8.4, D.8.5, D.8.6, D.9.3, D.9.4

7.5 Emerging Network Technologies

7.5.1 Standardization gaps on Emerging Network Technologies

The gaps included in this clause were extracted were extracted from the output of the Emerging Network Technologies gap analysis work included in Appendix V.

Gap E.1 Considering ICN as a protocol for IMT-2020 Network	Priority: High
Description: In the existing mobile infrastructure, IP is the main transport protocol and everything is optimized around the layer-3 OSI (TCP/IP) stack. However, experience has shown that there is a need to migrate to protocols which can comprehensively integrate infrastructure, transport, and content. Research and development in ICN shows the possibilities to solve this problem. ICN will need further development in areas such as mobility management, end-to-end QoS, prioritization and scale to manage billions of devices, which are framework of IMT-2020 networks. There are three possible scenarios for IMT-2020 network where ICN can be introduced Option 1: The IMT-2020 network using ICN as an overlay protocol Option 2: The IMT-2020 Network using IP as transport for mobility management and ICN for service delivery without an overlay. ICN routing exists on the UE and P-GW. Option 3: The IMT-2020 network using ICN as native transport for mobility management and service delivery. ICN routing exists throughout the network. Gap: Detailed architecture analysis of the three above options is required.	
Related work: IRTF/ICNRG documents	

Gap E.2 ICN – Robust header compression for air interface (PDCP)	Priority: High
Description: Applicable to Options 1, 2 and 3: Standard compression techniques, such as LZW, are not appropriate for ICN packets because many of the fields represent cryptographic octet strings and thus are not sub-string compressible. Gap:	

1. When encapsulated in IP, there is a need to specify an ICN profile, similar to an RTP profile. 2. When used as a native protocol, e.g. over an ICN slice, there is a need to specify an ICN--specific ROHC profile at the air interface.
Related work: IETF RFC 4995, 3GPP TS 36.300

Gap E.3 ICN – Mobility anchoring (ICN aware S-GW)	Priority: Medium
Description: Applicable to Option 3: Existing mobile networks have one mobility anchoring point (S-GW) so that devices can be located for downstream traffic. Each UE has one anchoring point and multiple service end points e.g. APN based upon simultaneous services being accessed by the application running device e.g. visual voice mail, VoLTE, mobile internet etc. For the IMT-2020 architecture, using ICN as a transport protocol, a change in the ICN specification is needed to support edge device anchoring. ICN allows for new mobility models not tied to the anchor-based approach used for IP. This is because ICN does not require a unique source representation (egress identity). It is necessary to modify and develop call flows for ICN based device attachment, authentication and registration with content providers. It is proposed to describe an ICN operating in three different models: 1. Similar to current single anchor like S-GW and P-GW 2. Using the closest ICN router(s) as a single attach 3. Distributed among points of attachment Gap: The 3GPP model and the ITU-R 5G document specify that a UE only has one S-GW, whereas for ICN, multiple simultaneous gateways would be required.	
Related work: ITU-R M.[IMT.ARCH], 3GPP TS23.401	

Gap E.4 ICN – Mobility (ICN-aware MME)	Priority: Medium
Description: Applicable to Option 3: In the existing LTE architecture, all mobility management is handled by the MME, eNodeB, etc. The ICN protocol must evolve to include mobility management messages. Mobility management can be done either by the MME or something similar e.g. ICN router(s)/edge gateway. The first step for introduction of ICN capabilities can be an ICN aware MME, S-GW/P-GW, etc. and eventually replacing the GTP based model with ICN based transport functions e.g. an ICN-aware MME should allow for one of the several ICN-style mobility models. Gaps: The 3GPP TS23.401 specification specifies mobility management and attachment procedures using IP. New specifications and procedures are necessary to use ICN as a transport protocol.	
Related work: ITU-R M.[IMT.ARCH], 3GPP TS23.401	

Gap E.5 ICN Protocol (ICN-aware P-GW operation)	Priority: Medium
Description: Applicable to Options 2 and 3: Description: Currently IP is used for UE attachment procedures for APN in the P-GW, which is the services attachment point. The P-GW manages IP address allocation, billing and policy enforcement etc. Gap: ICN mechanisms for P-GW functions such as billing, policy enforcement, etc. need to be specified.	
Related work: ITU-R M.[IMT.ARCH], 3GPP TS23.401	

Gap E.6 ICN Protocol Execution (slice)	Priority: High
Description: Applicable to Options 2 and 3: In the existing LTE architecture the main protocol is based on IP however for IMT-2020 there is a need to define how ICN protocols operate in the RAN and EPC. What computing, execution or hardware resources will be available? In the case of using a slice, there is a need to specifically enumerate the service interfaces and how those interfaces are exposed to non-IP based protocols operating within a slice environment. Gap: For a software-based slice environment, the ICN execution environment needs to be described, including the virtualized resources that are available and their interfaces?	
Related work: Network Softwareization	

Gap E.7 ICN – Lawful intercept (specify what to capture)	Priority: High
Description: Applicable to Options 2 and 3: In the existing LTE network, lawful intercept messages are based on IP and these messages are taken from gateways (SAEGW, MME, HSS etc.). ICN protocols may operate with a different model of SGW and PGW, such as those elements being collapsed to the base station (option 3). This may significantly affect how lawful intercept operates. As a non-IP protocol, additional collection practices may need to be specified and implemented for a specific ICN. When gateways are distributed, lawful intercept message have to be collected from multiple egress points. Gap: How to specify an intercept of a non-IP protocol, for example what does the packet filter look like?	
Related work:3GPP TS23.002	

Gap E.8 ICN mobility and routing	Priority: Medium

Description: Applicable to Options 2 and 3: There is a need to specify routing models for ICN within an IMT-2020 environment to enable desired "mobility" features. During initial ICN development work, mobility was not factored. However, the IMT-2020 network will have mobility and this has to be managed for millions of devices.

There are three possible scenarios for mobility

UE consumer mobility

UE producer mobility

ICN state transfer

Operator maintained content is cached for intra-RAT and inter-RAT data retrieval. For example, content with the same name may exist in multiple locations, but one does not want to create multi-homed routes.

Distributed routing within the RAN

Certain features, such as MEC, could benefit from local dynamic routing to solve the service rendezvous problem such that a UE application can easily exploit local services.

Disaster recovery or other edge applications could benefit from local dynamic routing.

Similar for IoT and M2M applications.

Gap: Study using the ICN routing and control for mobility management rather than the current anchoring based mechanisms.

Related work:3GPP TS23.401

Gap E.9 ICN UE provisioning	Priority: High

Description: Applicable to Options 2 and 3: ICN is a new technology that does not have the breadth of support that IP has for operations and management.

In the current mobile network, the UE identity (IP address) is allocated and managed by the P-GW based upon applications (APN). When ICN is used within the carrier network, then the carrier should be able to assign a name and other ICN parameters.

Gap: Define a protocol and mechanism for ICN provisioning.

Related work: 3GPP TS23.401

Gap E.10 ICN managing IMT-2020 Self Organizing Network (SON)	Priority: Medium

Description: Applicable to Option 3: SON needs to communicate between different radio, element management system (EMS), OSS/BSS and core network (CN). ICN doesn't support communication mechanisms for SON.

Gap: There is need to define the right set of ICN messages and parameters so that the SON platform is managed effectively.
Related work: 3GPP TS32.50X

Gap E.11 ICN – Operations and management (common interfaces)	Priority: Medium
Description: Applicable to Options 2 and 3: Current management platforms are defined to run over IP and manage IP. When ICN elements are introduced into the network the management protocol needs to support ICN. In Option 3, management protocols may need to operate over ICN. Gap: The IMT-2020 architecture describes common management interfaces that would need to be ICN aware.	
Related work: TMN documents	

Gap E.12 ICN – Operations and management (SDN/Openflow)	Priority: Medium
Description: Applicable to Options 2 and 3: Many of today's carrier networks are managed/programmed with SDN. Gap: Today's SDN tools need extensions for ICN.	
Related work: Openflow, ONOS, ODL, POF, P4	

Gap E.13 ICN – Security (authentication and encryption)	Priority: Medium
Description: Applicable to Options 2 and 3: Often each party in an ICN communication is considered to have a cryptographic identity. Should that identity be used in IMT-2020 associations or resource usage, or should it all be based on IMEI or existing IMT-2020 identity? Key resolution services: some ICN approaches use the idea of a 'key resolution service' to determine from a trusted anchor what public key a publisher should be using for its namespace. Is this needed in an IMT-2020 environment and if so how would this integrate in a carrier environment? Gap: IMT-2000/Advanced has defined several identities for a UE. IMT-2020 is also required to support several identities for a UE. A study is needed to determine if any of these identities are suitable for ICN, or if additional cryptographic identities are needed.	
Related work: 3GPP TS36.323, 3GPP TS33.401, 3GPP TS23.401	

Gap E.14 ICN – Security (encryption)	Priority: Medium

Description: Applicable to Options 2 and 3: Today, end-to-end encryption using IPSEC/TLS does not allow the carrier to intelligently manage the data (e.g. DPI, caching). ICN offers new possibilities for key exchange and encryption where the carrier can play an active role because the ICN packet can be selectively encrypted between the carrier and the remote party. Gap: ICN sometimes uses different forms of encryption than what is found in today's IP networks. IMT-2020 should study the use of selective ICN encryption.
Related work: None

Gap E.15 ICN – QoS (demand based)	Priority: Medium
Description: Applicable to Options 2 and 3: QoS is well defined in IP and Ethernet layer which is mapped to QoS Class Indicator (QCI). In Options 2 and 3, ICN can use a similar mapping of DSCP to support interoperability with existing transport networks. Additionally ICN provides the possibility for new forms of QoS definition. Gap: A study is needed on ICN-specific QoS in IMT-2020, for traffic prioritization and congestion management	
Related work: 3GPP TS23.401	

7.5.2 Recommendations to parent group on Emerging Network Technologies

The emerging network technologies group within FG-IMT-2020 has studied, in particular, the applicability of ICN to solve the IMT-2020 challenges. In Gap E.1 we have described three deployment options: (1) over-the-top, (2) continue using IP as transport for mobility management and ICN for service delivery without overlay, and (3) using ICN as native transport for mobility management and service delivery. Option (1) realizes no inherent benefits for IMT-2020 as it is purely over-the-top. Option (2) allows incremental introduction of ICN to carrier networks, and we view this as the most likely near-term use of ICN within IMT-2020. Option (3) requires further development of ICN and IMT-2020 standards to use ICN natively in place of IP transport.

ICN may offer technological benefits for realizing IMT-2020 goals. We recommend that the Study Group continue investigation of ICN and promote more in depth study of ICN inside IMT-2020 networks, particularly as a new service delivery platform that can unify the many different requirements of IoT, edge computing and content delivery.

Option (2) makes the use of ICN visible to the IMT-2020 network so it is not simply a bit-pipe for this new technology, but a participant. The use of ICN to replace IP as the transport protocol within IMT-2020 is still a stretch goal within the 2020 timeframe. While it is an interesting research area, we believe the current effort should go towards option (2).

Additional references related to ICN can be found in [b-ITU-T Y.3033] and [b-Y.supFNDAN];

8 Conclusion and future work

- Considering that the mandate (gap analysis) is successfully completed, the focus group should not continue in its current form;

- The focus group has identified a set of gaps that need to be addressed. An indication has been given where it should be done;
- Some aspects have not been addressed sufficiently due to time constraints, e.g., security and privacy;
- The major differences between 4G and 5G have been identified and considered during the preparation of the individual gaps.
- Some gaps do appear to be useful to continue moving forward with, considering the gaps, and in coordination with other SDOs:
  - Demonstrations or prototyping with other groups (e.g., with open source community);
  - Network softwarization and ICN;
  - Network architecture refinement;
  - Fixed mobile convergence;
  - Network slicing for front haul/back haul;
  - New traffic models and associated QoS and OAM aspects applicable to IMT-2020 architecture.

Bibliography

[b-ITU-T Y.3001] Recommendation (2012) - Future networks: Objectives and design goals - http://www.itu.int/rec/T-REC-Y.3001-201105-I;

[b-ITU-T Y.3021] Recommendation (2012) - Framework of Energy Saving in Future Networks - http://www.itu.int/rec/T-REC-Y.3021-201201-I/en;

[b-ITU-T Y.3031] Recommendation (2012) - Identification framework in future networks- http://www.itu.int/rec/T-REC-Y.3031-201205-I;

[b-ITU-T Y.3011] Recommendation ITU-T Y.3011 (2012), *Framework of network virtualization for future networks* www.itu.int/rec/T-REC-Y.3011-201201-I/en;

[b-ITU-T Y.3012] Recommendation ITU-T Y.3012 (2014), *Requirements of network virtualization for future networks* https://www.itu.int/rec/T-REC-Y.3012/en;

[b-ITU-T Y.3300] Recommendation (2014) - Framework of software-defined networking - https://www.itu.int/rec/T-REC-Y.3300-201406-I

[b-ITU-T Y.3033] Recommendation (2014) - Framework of data aware networking for future networks.

[b-Y.supFNDAN] Revised Draft of Y.supFNDAN – supplement to Y.3033 on scenarios and use cases of data aware networking (July 13-25, 2015, Geneva).

[b-ITU-T Y.3500] Recommendation (2014) |– Cloud computing – Overview and vocabulary- http://www.itu.int/rec/T-REC-Y.3500-201408-I

[b-ITU-T Y.3021] Recommendation (2013) - Cloud computing infrastructure requirements - ww.itu.int/ITU-T/recommendations/rec.aspx?rec=11918

[b-ITU-T Y.3502] Recommendation (2014) - Cloud computing - Reference architecture www.itu.int/ITU-T/recommendations/rec.aspx?rec=12209

[b-ITU-T Y.3511] Recommendation (2014) - Framework of inter-cloud computing - https://www.itu.int/rec/T-REC-Y.3511/en

[b-ITU-T Y.3512] Recommendation (2014) - Cloud computing - Functional requirements of Network as a Service- http://www.itu.int/rec/T-REC-Y.3512-201408-P

[b-ITU-T Y.3513] Recommendation (2014) - Cloud computing - Functional requirements of Infrastructure as a Service - http://www.itu.int/rec/T-REC-Y.3513-201408-I

[b-ITU-T Y.1540] Recommendation (2011) - Internet protocol data communication service - IP packet transfer and availability performance parameter - http://www.itu.int/rec/T-REC-Y.1540-201103-I/en;

[b-ITU-T Y.1541] Recommendation (2011) - Network performance objectives for IP-based services - http://www.itu.int/rec/T-REC-Y.1541-201112-I/en;

[b-ITU-T Y.1542] Recommendation (2010) - Framework for achieving end-to-end IP performance objectives - http://www.itu.int/rec/T-REC-Y.1542-201006-I/en;

[b-ITU-T G.1000] Recommendation (2001) - Communications Quality of Service: A framework and definitions - http://www.itu.int/rec/T-REC-G.1000-200111-I/en

[b-ITU-T G.1010] Recommendation (2001) - End-user multimedia QoS categories - http://www.itu.int/rec/T-REC-G.1010-200111-I/en;

[b-ITU-R M.2083-0] Recommendation ITU-R M.2083-0 (2015), Framework and overall objectives of the future development of IMT for 2020 and beyond.

https://www.itu.int/rec/R-REC-M.2083

[b-ITU-R M.2375-0] Reports ITU-R M.2375-0, Architecture and topology of IMT networks https://www.itu.int/pub/R-REP-M.2375-2015

[b-ETSI MEC] Draft ETSI GS MEC 002, Mobile-Edge Computing(MEC);Technical Requirements v.0.4.2 (2015-07-30)

[b-ETSI MEC] Draft ETSI GS MEC 003, Mobile Edge Computing(MEC);Framework and reference architecture v0.0.1 (2015-06-05)

[b-3GPP TS 23.107] Technical Specification (2014) - Quality of Service (QoS) concept and architecture - http://www.3gpp.org/ftp/specs/archive/23_series/23.107/

[b-3GPP TS 23.401] Technical Specification (2015) - General Packet Radio Service (GPRS) enhancements for Evolved Universal Terrestrial Radio Access Network (E-UTRAN) access - http://www.3gpp.org/ftp/specs/archive/23_series/23.401/

[b-ETSI NFV] ETSI NFV ISG, Network Functions Virtualisation, http://portal.etsi.org/portal/server.pt/community/NFV

[b-IETF RFC 3746] IETF RFC 3746 (2004), Forwarding and Control Element Separation (ForCES) Framework.

[b-IETF RFC 7149] IETF RFC 7149 (2014), Software-Defined Networking: A Perspective from within a Service Provider Environment.

[b-IETF SFC] IETF Service Function Chaining (sfc) working group, http://datatracker.ietf.org/wg/sfc/charter/

[b-ONF] Open Networking Foundation, "OpenFlow/Software-Defined Networking (SDN)," https://www.opennetworking.org/.

[b-SDN-WS Nakao] "Deeply Programmable Network", Emerging Technologies for Network Virtualization, and Software Defined Network (SDN), ITU-T Workshop on Software Defined Networking (SDN), http://www.itu.int/en/ITU-T/Workshops-and-Seminars/sdn/201306/Documents/KS-Aki_Nako_rev1.pdf.

[b-Programmable Networks - Galis]–"Programmable Networks for IP Service Deployment" ISBN 1-58053-745-6, pp450, June 2004, Artech House Books, http://www.artechhouse.com/International/Books/Programmable-Networks-for-IP-Service-Deployment-1017.aspx,

[b-Cloud Survey –Heilig] - "A Scientometric Analysis of Cloud Computing Literature"- a review of approx. 25,000 papers] - IEEE Transactions on Cloud Computing, Volume: PP, Issue: 99, 30 April 2014, ISSN: 2168-7161; DOI: 10.1109/TCC.2014.2321168

[b-ETSI NFV MANO] Network Functions Virtualisation (NFV) Management and Orchestration- http://www.etsi.org/deliver/etsi_gs/NFV-MAN/001_099/001/01.01.01_60/gs_nfv-man001v010101p.pdf

[b-FG IMT-2020 WS 5GMF Nakao] – Akihiro Nakao, "Overview of network softwarization and adoption to 5G", ITU-T Focus Group on IMT-2020, Pre-meeting workshop on network, softwarization, Turin, Italy, 21 September 2015

[b-FG IMT-2020 WS Galis] – Alex Galis, "Challenges in network softwarization", ITU-T Focus Group on IMT-2020, Pre-meeting workshop on network softwarization, Turin, Italy, 21 September 2015

[b-FG IMT-2020 WS Manzalini] - Antonio Manzalini, Telecom Italia / IEEE SDN Committee: "R&D status of network softwarization", ITU-T Focus Group on IMT-2020, Pre-meeting workshop on network softwarization, Turin, Italy, 21 September 2015

[b-FG IMT-2020 WS Wang] Yachen Wang, China Mobile: "Key technologies to support network softwarization", ITU-T Focus Group on IMT-2020, Pre-meeting workshop on network softwarization, Turin, Italy, 21 September 2015

[b-FG IMT-2020 WS Tsuda] Toshitaka Tsuda, Waseda University: "CCN implementation by network softwarization", ITU-T Focus Group on IMT-2020, Pre-meeting workshop on network softwarization, Turin, Italy, 21 September 2015

[b-4G America] *5G technology evolution recommendations*

[b-NGMN] *NGMN 5G white paper*

[b-METIS] *Updated scenarios, requirements and KPIs for 5G mobile and wireless system with recommendations for future investigations*

[b-MEF 22.1.1] *Mobile back haul Implementation Agreement*

[b-DMTF OVF] Distributed Management Task Force (DMTF) Open Virtualisation Format (OVF) specification document number DSP0243_1.1.0, 2010-01-12 http://www.dmtf.org/sites/default/files/standards/documents/DSP0243_1.1.0.pdf

[b-P4] P4: Pat Bosshart, Dan Daly, Martin Izzard, Nick McKeown, Jennifer Rexford, Cole Schlesinger, Dan Talayco, Amin Vahdat, George Varghese, David Walker, "Programming Protocol-Independent Packet Processors", http://arxiv.org/abs/1312.1719

[b-FLARE] FLARE: Nakao, Akihiro. "Software-defined data plane enhancing SDN and NFV." IEICE Transactions on Communications 98.1 (2015): 12-19.

Appendix I

High-level network architecture for IMT-2020

Editor's Note: Appendix I was produced during the FG-IMT 2020 focus group in order to investigate gaps in standardization related to IMT-2020. While the request from SG-13 was to deliver a report outlining standardization gaps, the consensus of the focus group was that the working documents produced and used during the focus group work contained useful information for future work and should be captured. Note, however, the focus group concentrated on producing accurate descriptions of the standardization gaps in the main body of this document; some minor errors may exist in the appendices. They are, however, the output of the focus group but are provided for information only.

Editor's Note: This appendix uses clause references in a form usually associated for normative text. This is maintained for this report to align with references made in the main body of this report.

This is the final output document for the high-level network architecture for IMT-2020, which will be integrated into a report to be submitted to SG13 plenary as a result of Focus Group on IMT-2020.

This work has been done with the help of many contributors from various organizations. We appreciate all the contributors for the active works to improve this document.

1 Scope

This document describes the high-level view of network architecture for IMT-2020 including requirements, gap analyses, and design principles of IMT-2020 with the aim of giving directions to the relevant Study Groups in ITU-T in developing standards on network architecture in IMT-2020. It should be noted that this document is based on the related works in ITU-R and other SDOs.

2 References

The following ITU-T Recommendations and other references contain provisions, which, through reference in this text, constitute provisions of this Recommendation. At the time of publication, the editors indicated were valid. All Recommendations and other references are subject to revision; users of this Recommendation are therefore encouraged to investigate the possibility of applying the most recent edition of the Recommendations and other references listed below. A list of the currently valid ITU-T Recommendations is published regularly.

Note – The reference to a document within this Recommendation does not give it, as a stand-alone document, the status of a Recommendation.

- 4G America, *5G technology evolution recommendations*.
- ITU-T Recommendation Y.2060, *Overview of the Internet of things*.

- ITU-T Recommendation Y.2065, *Service and capability requirements for e-health monitoring services.*
- ITU-R Recommendation M.2083-0, *Framework and overall objectives of the future development of IMT for 2020 and beyond.*
- METIS, *Updated scenarios, requirements and KPIs for 5G mobile and wireless system with recommendations for future investigations.*
- NGMN, *NGMN 5G white paper*.

3 Terms and definitions

This deliverable defines the following terms:

Peak data rate is maximum achievable data rate under ideal conditions per user/device (in Gbit/s).

User experienced data rate is achievable data rate that is available ubiquitously[4] across the coverage area to a mobile user/device (in Mbit/s or Gbit/s).

Latency is the contribution by the network to the time from when the source sends a packet to when the destination receives it (in ms).

Mobility in performance target point of view is the maximum speed at which a defined QoS and seamless transfer between radio nodes which may belong to different layers and/or radio access technologies (multi-layer/-RAT) can be achieved (in km/h).

Connection density is a total number of connected and/or accessible devices per unit area (per km^2).

Energy efficiency is energy efficiency has two aspects:

- on the network side, energy efficiency refers to the quantity of information bits transmitted to/ received from users, per unit of energy consumption of the radio access network (RAN) (in bit/Joule);
- on the device side, energy efficiency refers to quantity of information bits per unit of energy consumption of the communication module (in bit/Joule).

Spectrum efficiency is average data throughput per unit of spectrum resource and per cell[5] (bit/s/Hz).

Area traffic capacity is total traffic throughput served per geographic area (in Mbit/s/m^2).

4 Abbreviations and acronyms

This deliverables defines the following abbreviations:

AAA	Authentication, Authorization, Accounting
APN	Application Network
AR	Augmented Reality
CGF	Converged Gateway Function

4 The term “ubiquitous” is related to the considered target coverage area and is not intended to relate to an entire region or country.

5 The radio coverage area over which a mobile terminal can maintain a connection with one or more units of radio equipment located within that area. For an individual base station, this is the radio coverage area of the base station or of a subsystem (e.g. sector antenna).

EPC	Evolved Packet Core
GW	Gateway
ICN	Information Centric Networking
IMT	International Mobile Telecommunications
IoT	Internet of Things
IP	Internet Protocol
KPI	Key Performance Index
LWA	LTE/WiFi Link Aggregation
NFV	Network Function Virtualization
MNO	Mobile Network Operator
MTC	Machine Type Communication
NAS	Non-Access Stratum
PGW	Packet Data Network Gateway
QoE	Quality of Experience
QoS	Quality of Service
RAT	Radio Access Technology
SDN	Software Defined Networking
TWAG	Trusted Wireless Access Gateway
UCF	Unified Control Function
UE	User Equipment
UHD	Ultra High Definition
VNF	Virtual Network Function
UE	User Equipment
UHD	Ultra High Definition
VNF	Virtual Network Function

5 IMT-2020 use cases

In this clause, we review several use cases among others where IMT-2020 is used to enhance the services and/or IMT-2020 should be enhanced to support the services.

5.1 Smart Grid

The Smart Grid is a new electricity network, which highly integrates the advanced sensing and measurement technologies, information and communication technologies (ICTs), analytical and decision-making technologies, automatic control technologies with energy and power technologies and infrastructure of electricity grids.

The followings are the features which smart grid should support and also IMT-2020 has to satisfy to provide Smart Grid services on the network.

– *Observability*: It enables the status of electricity grid to be observed accurately and timely by using advanced sensing and measuring technologies

– *Controllability*: It enables the effective control of the power system by observing the status of the electricity grid

– *Timely analysis and decision-making*: It enables the improvement of intelligent decision-making process

– *Self-adapting and self-healing capability*: It prevents power disturbance and breakdown via self-diagnosis and fault location

– *Integration of renewable energy integration*: It enables integrating the renewable energy such as solar and wind as well as the electricity from micro-grid to support efficient and stable energy delivery services for electric vehicle, smart home and others

5.2 E-Health

E-Health (electronic health) can be defined as the cost-effective and secure use of information and communications technologies in support of health and health-related fields including health-care services, health surveillance, health literature, health education, health knowledge, and health research. E-Health systems continue to hold great promise for improving global access to health-care services and health informatics, particularly in the developing world. The advancements of E-Health in remotely administered medicine will also increasingly enable virtual multimedia delivery of medical consultation, remote imaging services, specialized medical diagnostics, remote medical procedures, etc. Standardized electronic medical records promise to facilitate the digital exchange of patient data among a patient's primary care physician and other health providers.

The followings are the features which E-Health should support and also IMT-2020 has to satisfy to provide E-Health services on the network.

– *Various access and massive communication*: The increasing use of diagnostic tools such as 3D and 4D ultrasounds, CAT scans and MRIs, and the miniaturization of this equipment to a portable/hand-held form factor will lead to even higher demands being placed on wireless networks. In addition, Bio-connectivity, which is the continuous and automatic medical telemetry (e.g., temperature, blood pressure, heart-rate, blood glucose) collection via wearable sensors, is another strong emerging trend that will add to the wireless communications requirements.

– *Robust infrastructure*: Electronic health records is still a central focus of standards organizations, national e-health policies, and strategies within health systems. This is also an area that continues to raise concerns about data security and privacy. As the use of massive distributed computing resources to assist development of disease diagnostic and prevention progresses are heard for standards that provide for robust infrastructure, deployment models and interfaces.

– *Real time multimedia interactions*: The use of telecommunications networks and information technology for healthcare services such as remote clinical care, diagnostics, and electronic patient monitoring. Developments in this area include advancements in remote clinical care technologies that enable doctors to provide medical assessments and treatments from a remote location away from the patient via real time multimedia interactions with a patient such as a video feed transmitted over a telecommunications network.

– *Ultra-reliable and low latency communications*: The network provider needs support for providing access to E-Health applications as fast as possible upon service request.

– *Quality of service*: The network provider needs support for obtaining the customer's E-Health service-related information in order to allocate or configure for the E-Health customer the appropriate network resources, such as IP address, network bandwidth, QoS policy, and so on.

– *Policy-based communication*: The network layer is required to provide policy-based communication for E-Health applications and E-Health devices. Policy is a set of rules whose variables include, but are not limited to, time, bandwidth, data throughput, network type, traffic priority, and so on. By means of policy-based communication, E-Health applications and E-Health devices can obtain the desired QoS.

– *Network-based locating*: The network layer is recommended to provide the location related information from the network layer (e.g., IP address, access point location, and so on) for locating the position of E-Health devices. Event triggered location information notification is recommended to be supported.

– *Network resource provision*: The network layer is required to provide the network resource provision capability for E-Health applications and E-Health devices. Depending on the specific deployment of E-Health applications and E-Health devices, E-Health applications and E-Health devices may automatically use these provided network resources and configure themselves to connect to the network directly. In this way, E-Health customers can use the E-Health services directly, without the need to configure the E-Health devices

– *Secure communications*: The information carried by the E-Health services may be delivered across different administrative domains (e.g., countries, operators). The E-Health system supports secure communications between different domains. The information exchanged between different domains must be protected from random errors, as well as snooping or hacking attacks.

– *Confidentiality*: Whenever information is exchanged, stored or processed, the confidentiality of the data must be enforced and safeguarded by the E-Health system. All exchanges of data between e-health partners, for example E-Health device provider, E-Health application provider, network provider and platform provider, must be performed in a way that prohibits any unwanted disclosure of data, e.g., to third parties.

– *Integrity*: The integrity of the transmitted information must be guaranteed; transmitted data from the sender should be received without any alteration. It must be identified that the transmitted data have not been damaged, reduced or altered. Any loss of integrity of the transmitted data must be recognizable by the recipient.

– *Data storage security*: It is recommended to support data storage security strategies including, but not limited to, data backup, anti-hacker data protection, uninterruptible power of data storage, data integrity validation and data recovery. In addition, data access control is required to be supported for privacy

5.3 Autonomous car

An autonomous car, also known as an uncrewed vehicle, driverless car, self-driving car and robotic car, is an autonomous vehicle which is capable of fulfilling the main transportation capabilities of a traditional car by sensing its environment and navigating without human input. The implementation of autonomous cars could theoretically lead to many improvements in transportation, including a reduction in car accidents and obstacles successfully.

For this, V2X techniques may be used. Connected-vehicle systems use wireless technologies to communicate in real time from vehicle to vehicle (V2V) and from vehicle to infrastructure (V2I), and

vice versa. The convergence of communication- and sensor-based technologies, therefore, could deliver better safety, mobility, and self- driving capability than either approach could deliver on its own.

The followings are the features which autonomous car should support and also IMT-2020 has to satisfy to provide autonomous car services on the network.

- *Dependency on Sensors*: Although connected vehicle solutions can communicate with the external environment, sensor-based solutions will need to co-exist in order to cover situations that involve obstacles − for example, obstructions in the road or pedestrians− that would not be connected and communicating with the network.
- *Data Security*: Numerous security threats will arise once personal mobility is dominated by self-driving vehicles. Unauthorized parties, hackers, or even terrorists could capture data, alter records, instigate attacks on systems, compromise driver privacy by tracking individual vehicles, or identify residences.
- *Enhanced massive vehicle type communications*: A significant number of connected devices are expected to use IMT systems. In addition, as more and more things get connected, various services that utilize the connection capabilities of things will appear.
- *Ultra-reliable and low latency communications*: Reliability and latency are the essentially required for the safe running of autonomous cars.

5.4 Internet of things

A global infrastructure for the information society, enabling advanced services by interconnecting (physical and virtual) things based on existing and evolving interoperable information and communication technologies. From the perspective of technical standardization, the IoT can be viewed as a global infrastructure for the information society, enabling advanced services by interconnecting (physical and virtual) things based on existing and evolving interoperable information and communication technologies (ICTs). Physical things exist in the physical world and are capable of being sensed, actuated and connected. The examples of physical things include the surrounding environment, industrial robots, goods and electrical equipment. Virtual things exist in the information world and are capable of being stored, processed and accessed. The examples of virtual things include multimedia content and application software. Through the exploitation of identification, data capture, processing, and communication capabilities, the IoT makes full use of things to offer services to all kinds of applications, whilst ensuring that security and privacy requirements are fulfilled.

The followings are the features which IoT should support and also IMT-2020 has to satisfy to provide IoT services on the network.

- *High connection density*: With regard to the IoT, anything can be interconnected with the global information and communication infrastructure.
- *Multiple heterogeneous access networks*: The devices in the IoT are heterogeneous as based on different hardware platforms and networks. They can interact with other devices or service platforms through different networks
- *Autonomic networking*: Autonomic networking (including self-management, self-configuring, self-healing, self-optimizing and self-protecting techniques and/or mechanisms) needs to be supported in the networking control functions of the IoT, in order to adapt to different application domains, different communication environments and large numbers and types of devices.

– *Security*: In the IoT, every 'thing' is connected which results in significant security threats, such as threats towards confidentiality, authenticity and integrity of both data and services. A critical example of security requirements is the need to integrate different security policies and techniques related to the variety of devices and user networks in the IoT.

– *Manageability*: Manageability needs to be supported in the IoT in order to ensure normal network operations. IoT applications usually work automatically without the participation of people, but their whole operation process should be manageable by the relevant parties.

– *Energy efficiency*: Sensor needs long life time activities.

– *Networking capabilities*: Providing relevant control functions of network connectivity, such as access and transport resource control functions, mobility management or authentication, authorization and accounting (AAA).

– *Transport capabilities*: Focus on providing connectivity for the transport of IoT service and application specific data information, as well as the transport of IoT-related control and management information.

– *Local network topology management*: Traffic and congestion management, such as the detection of network overflow conditions and the implementation of resource reservation for time-critical and/or life-critical data flows

6 Performance targets for IMT-2020

The performance parameters for so-called 5G network are defined separately in related SDOs, research projects, and industry associations; the parameters are similar but the target values of the parameters show subtle differences with each other. ITU-R, among them, identified the following eight key capabilities and the targets for IMT-2020 in recommendation M.2083-0 , mostly from the radio access network points of view while they are still quite relevant to its network as well, which are subject to be changed according to future studies.

Table 1. IMT-2020 key capabilities in ITU-R

Parameters	Target
User experienced data rates	100 Mbps
Peak data rates	20 Gbps
Mobility	up to 500 km/h with acceptable QoS
Latency (air interface)	1 ms
Connection density	10^6 /km^2
Network energy efficiency	100 times better than IMT-Advanced
Spectrum efficiency	3 times better IMT-Advanced
Area traffic capacity	10 Mbit/s/m^2

All the other performance targets from ITU-R can be directly applied to the design of network architecture, but the target for the latency is set only for the air interface in M.2083-0. However, when we consider the key capabilities from other sources such as METIS 2020 project, the target for the end-to-end latency is expected to be around 5 ms ~ 10 ms depending on the service characteristics.

7 Requirements and gap analysis

7.1 Enhanced mobile broadband services

More and more user devices are being equipped with enhanced media consumption capabilities, such as Ultra-High Definition display, multi-view High Definition display, mobile 3D projections, immersive video conferencing, and augmented reality and mixed reality display and interface. This will all lead to a demand for significantly higher data rates in IMT-2020.

The demand for mobile high-definition multimedia also keeps increasing in many areas beyond entertainment, such as medical treatment, safety, and security, which is well reflected to the performance targets for connection density and area traffic capacity in ITU-R recommendation M.2083-0.

Editor's note: the tables containing the descriptions of the gaps have been deleted from this appendix. The final texts describing the standardization gaps are included in the main body of this Report. Where a draft table existed, a reference to the specific text in Clause 7 of the report is made.

Gap A.1: Various bandwidth/data-rates demands. See Clause 7.1.1 of the main body of this report.

Gap A.2: Complex connectivity model. See Clause 7.1.1 of the main body of this report.

Gap A.3: Application-aware and distributed network architecture. See Clause 7.1.1 of the main body of this report.

7.2 Enhanced massive machine type communications

In IMT-2020 networks, almost every object that can benefit from being connected is expected to be connected through wired or wireless internet technologies, which will lead to a situation where the number of connected devices exceeds the number of human user devices. These connected "things" can be various ranging from low-complexity devices to highly complex and advanced devices. As more and more things get connected, various services that utilize the connection capabilities of things will appear: smart energy distribution grid system, agriculture, healthcare, vehicle-to-vehicle and vehicle-to-road infrastructure communication.

At least one hundred thousand simultaneous active connections per square kilometre, which will be mostly coming true by the deployment of those massive MTC (machine type communications) devices, shall be supported in an IMT-2020 network. Consistent end-to-end user experience should also be provided even in the presence of that large number of concurrent connections.

Gap A.4: Signalling complexity in massive MTC. See Clause 7.1.1 of the main body of this report.

7.3 Ultra-reliable and low latency communications

The new applications with very low latency and real-time constraints are expected to be prevalent in IMT-2020 networks: driverless cars, enhanced mobile cloud services, real-time traffic control optimization, emergency and disaster response, smart grid, e-health, augmented reality, remote tactile control, and tele-protection are some of the examples.

Gap A.5: Increasing service availability. See Clause 7.1.1 of the main body of this report.

Gap A.6: Signalling to reduce end-to-end complexity. See Clause 7.1.1 of the main body of this report.

Gap A.7: End-to-end network latency model. See Clause 7.1.1 of the main body of this report.

7.4 Flexibility and programmability

An IMT-2020 network, as an integrated common core network, will be flexible enough to support extremely variety of requirements in user devices and application services. Therefore, the IMT-202 network is envisioned as a network where multiple logical network instances tailored to various requirements can be created. As a basic feature to realize this, the separation of control and data planes in IMT-2020 network is needed, which enables the components of an IMT-2020 network to be reconfigured, upgraded or even replaced easily with those of other vendors. NFV is expected to do a significant role in making the IMT-2020 network more flexible by realizing network components as software components. We should note that the reality would not allow all the required network functions to be softwarized mainly because of the performance reason.

The openness given by the separation of control and data planes also makes the network programmable by controlling/steering traffic depending on user specific requirements and some use-cases.

Gap A.8: Mobile network optimized softwarization architecture. See Clause 7.1.1 of the main body of this report.

Gap A.9: Data plane programmability. See Clause 7.1.1 of the main body of this report.

7.5 Quality of service

IMT-2020 network should provide consistent user experience and differentiated services in various aspects such as throughput, latency, resilience and costs per bit depending on service level agreement (SLA) of a user or its application.

Gap A.10: End-to-end QoS framework. See Clause 7.1.1 of the main body of this report.

7.6 Energy efficiency

IMT-2020 networks should meet all the other requirements and challenges in energy efficient manners. An IMT-2020 network should support up to 100 times better energy efficiency than IMT-Advanced in spite of 1000 times traffic increase without sacrificing the other performance targets. In reality, however, appropriate trade-off will be allowed between the energy efficiency and other performance requirements depending on the characteristics of user devices or applications.

Gap A.11: Energy efficiency. See Clause 7.1.1 of the main body of this report.

7.7 Enhanced privacy and security

IMT-2020 networks should provide robust and secure solutions for mission-critical applications such as smart girds, telemedicine, public safety, etc. to counter the threats to security and privacy brought by new radio technologies, new services and new deployment cases.

Gap A.12: Enhancement of privacy and security. See Clause 7.1.1 of the main body of this report.

Gap A.13: Enhancement identity management. See Clause 7.1.1 of the main body of this report.

7.8 Multiple heterogeneous radio access networks

The use of multiple heterogeneous radio access networks, including Wi-Fi networks, and their interworking with each other in existing IMT networks are becoming prevalent with various approaches: LTE/WiFi link Aggregation (LWA), interworking with ePDG or TWAG (trusted wireless access gateway), etc. The trend is also expected to be continued in IMT-2020 networks, but with more advanced and efficient ways. The multi-connectivity through the multiple available radio access networks improves the robustness of the network as well as the throughput performance. Especially, the dual connectivity of an existing IMT-network and a new IMT-2020 network will help ensure the smooth introduction of IMT-2020 networks.

Gap A.14: Multi-RAT connectivity. See Clause: 7.1.1 of the main body of this report.

7.9 Fixed-mobile convergence

IMT-2020 core network is envisioned as a converged access-agnostic core (i.e., where identity, mobility, security, etc. are decoupled from the access technology), which integrates multiple heterogeneous radio access networks as well as fixed networks on an IP basis.

Gap A.15: Fixed mobile convergence. See Clause 7.1.1 of the main body of this report.

Enhanced on-demand mobility management

IMT-2020 should support a wide range of mobility options. Only around 30 % of total subscribers are actually mobile even in the current IMT networks according to some of MNOs's observation. Therefore, IMT-2020 should not assume the same mobility support for all devices and application services but rather provide mobility on demand only to those that need it. While mobility is not required for some stationary devices such as smart meters and CPE devices, but we also need to provide mobility for high-speed trains running at 500 km/h. The service continuity levels also varies: seamless mobility, nomadic mobility, mobility for sporadic transmission, etc.

Gap A.16: Flexible mobility. See Clause 7.1.1 of the main body of this report.

Gap A.17: Mobility management for distributed flat network. See Clause 7.1.1 of the main body of this report.

7.10 Operation and management

The current status of having its own proprietary network management protocol in every different vendor should be enhanced by supporting a standard open interface so that a convergence network management system can control all different network devices as shown Figure 1. In addition to the standard management protocol, an IMT-2020 network should support common OAM protocols in IP-based transport network to enhance sustainability and reliability.

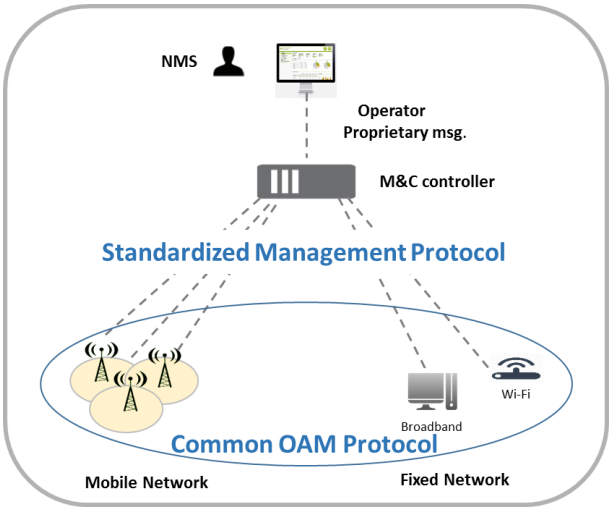


Figure 1. Standardized network management and OAM.

Gap A.18: End-to-end network management in a multi-domain environment. See Clause 7.1.1 of the main body of this report.

Gap A.19: OAM protocols. See Clause 7.1.1 of the main body of this report.

8 Structuring of IMT-2020 requirements

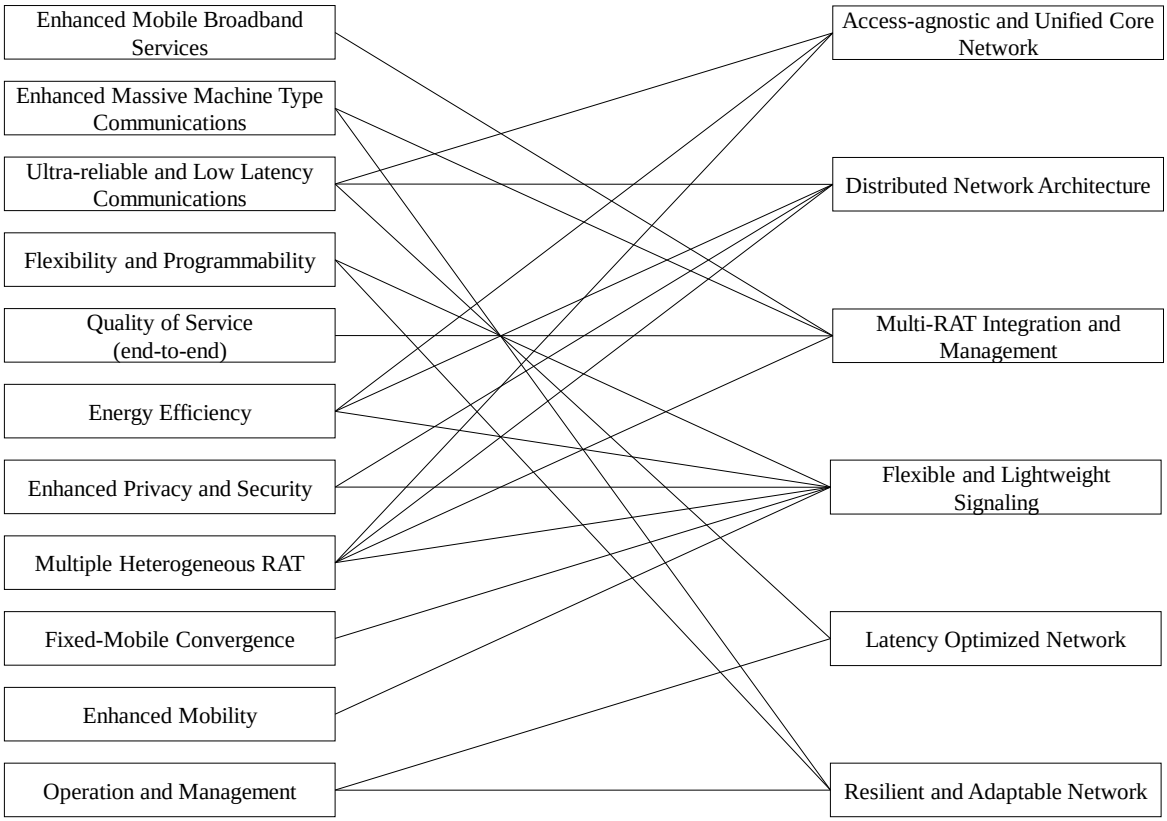


Figure 2. Structuring of requirements.

Figure 2 describes the structuring IMT-2020 requirements based on the gap analysis to give an overview of vision of IMT-2020 network architecture.

8.1 Access network-agnostic and unified core network

The introduction of a new mobile technology has been accompanied with a new packet core network in the legacy IMT networks. Therefore, the interworking issue between the new core network and legacy core networks has always been a technical challenge to overcome.

The IMT-2020 network is envisioned to be an access network-agnostic architecture whose core network will be a common unified core network for emerging new radio access technologies for IMT-2020 as well as existing fixed and wireless networks (e.g., Wi-Fi). The access technology-agnostic unified core network should be accompanied by common control mechanisms which are decoupled from access technologies.

8.2 Distributed network architecture

The IMT-2020 network should be flexible enough to handle the explosive increase of traffic from the new emerging bandwidth-hungry services such as ultra-high definition (UHD) TV, augmented reality (AR), video conferencing, remote medical treatment, etc. The heavily centralized architecture onto an anchor node of existing IMT networks is expected to be changed to cope with the explosion of mobile data traffic. This will require the gateways to the core network are expected to be located closer to the cell sites resulting in a distributed network architecture.

The distributed network architecture will bring a significant reduction on backhaul and core network traffic by enabling placing content servers closer to mobile devices and also be beneficial to the latency of the services.

8.3 Integrated management of multi-RAT and fixed access networks

The IMT-2020 network should facilitate an integrated access network management for multiple radio access networks including Wi-Fi and fixed access networks. The integrated multi-RAT and fixed network management should support seamless and consistent user experience while moving across different access networks, and also steer mobile devices to choose the most suitable access technology in a seamless way. In addition, simultaneous multiple connections for a mobile device to multiple RATs and fixed access networks should be supported to increase user experienced data rate through the integrated management of multi-RAT and fixed access networks.

8.4 Flexible signaling

In the existing IMT networks, all different types of traffic are treated in a uniform procedure by a monolithic bearer management and its accompanied signaling protocol. The characteristics of traffic are expected to vary significantly from devices to devices and from applications to applications in IMT-2020 networks. For example, as the number of IoT devices is expected to increase, the intermittent short burst traffic from massive number of devices will cause signaling bottleneck. Meanwhile, a full-fledged signaling may be still essential for the devices/applications in which the support of mobility is more critical. Therefore, the IMT-2020 network should support flexible signaling architecture.

8.5 Latency optimized network

The IMT-2020 network should be based on a proper network architecture to support low-latency network services. The end-to-end latency may come along with the support of distributed network architecture where application servers placed on network edges and flexible signaling architecture in which light weight signaling can be supported for latency critical applications.

8.6 Extensible network

The IMT-2020 network should be flexible and extensible to cope with various and sometimes conflicting service requirements adaptably instead of having a dedicated network for each emerging new service. The IMT-2020 network should be future-proof as much as possible to accommodate even unforeseeable use cases. The clear separation of control and data planes and the enabling technologies are the basis to make the IMT-2020 network flexible and extensible.

9 High-level IMT-2020 network architecture

The IMT-2020 network architecture should be defined in the end as a reference architecture which defines the functional entities and the interaction flows among the entities as well. In this document, however, we only define a high-level view of IMT-2020 network architecture as a guide for the reference architecture in the future standardization phase.

Network softwarization definitely will be one of the key technologies for IMT-2020 network to cope with the various service requirements. The detailed softwarized IMT-2020 network is handled in separate chapter. In this chapter, therefore, we focus on more fundamental elements which will address the other requirements of the IMT-2020 network.

9.1 IMT-2020 network architecture

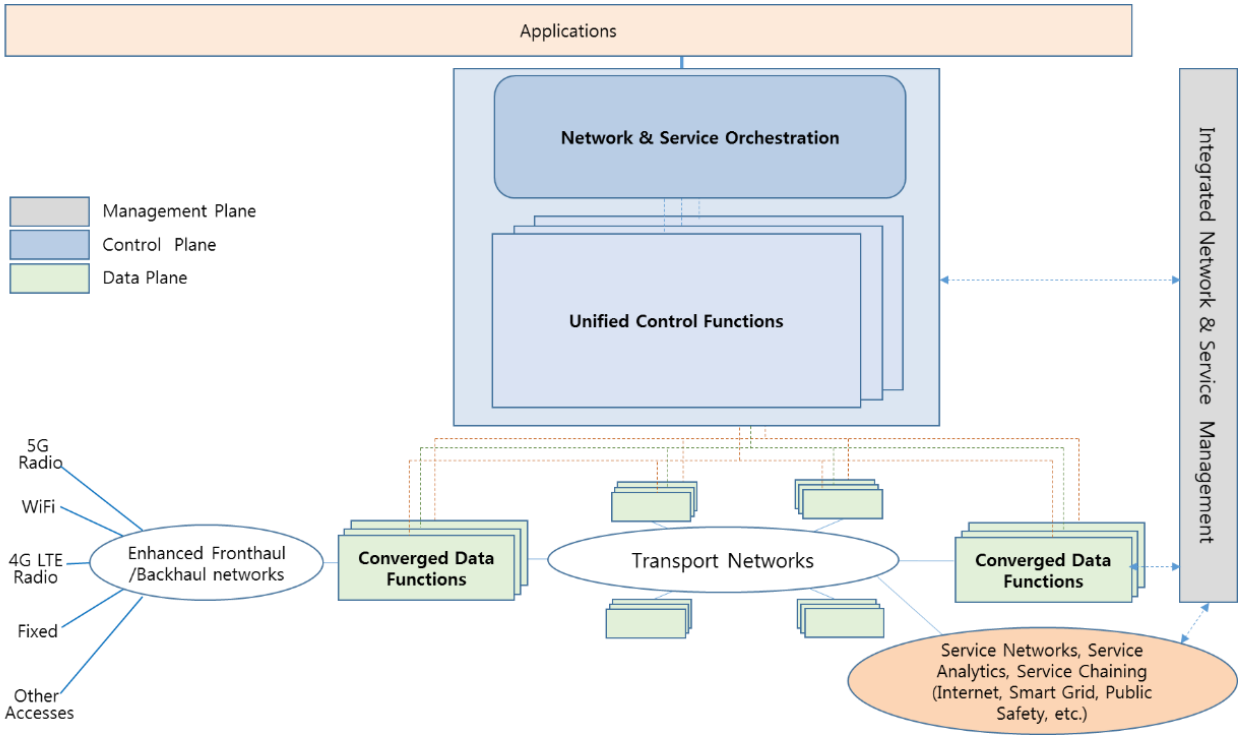


Figure 3. High-level IMT-2020 network architecture.

Error! Reference source not found. shows the high-level IMT-2020 network architecture. An IMT-2020 network differentiates itself from legacy IMT networks by further evolution and revolution as well in radio network, front/back-haul networks, and core networks. Multiple various access points, including a new IMT-2020 RATs, Wi-Fi AP, and even fixed networks, are connected to a converged data plane functions via an integrated access network so that mobile devices can be serviced through an access technology-agnostic network core. The converged data plane functions are distributed to the edges of an IMT-2020 common core network resulting in creating a distributed flat network. The control plane functions, which are responsible for mobility management, QoS control, etc., controls the user traffic to be served agnostically to the access networks to which it is attached.

IMT-2020 network architecture also can support massive native flexibility with the support of network softwarization (network virtualization, functions virtualization, programmability, etc.) depending on different service scenarios and requirements, which is a major differentiation from existing IMT networks.

9.2 Functional elements in IMT-2020 network architecture

The key parts among all the required functional elements are described here. The detailed functional architecture and definition of reference points will be discussed in standardization phase.

9.2.1 Data plane functions

The key data plane functions, which is expressed as converged data functions in the **Error! Reference source not found.**, include the following.

- IP flow management and control function:
  This function provides a method to classify user traffic into individual service flows which can be added, modified, and deleted separately. Quality of service is managed and guaranteed per service flow.

- Application identification function:
  This function is a traffic identification function to help control application flows according to their service characteristics in data plane.
- Multi-RAT coordination function:
  This function monitors the condition of multiple wireless networks and control the traffic flows among selected wireless networks. The most suitable RAT may be selected as a sole access network or multiple RATs can be selected to increase user experienced data rate.
- Computing and storage function:
  This function provides the computing and storage resources at the edges of IMT-2020 core network to support low latency services such as tactile internet more efficiently.

9.2.2 Control plane functions

The control plane functions may be centralized while parts of them may be locally distributed depending on the requirements and also for the purpose of efficiency, each of which is expressed as core control functions and access control functions, respectively.

The key control plane functions include the following.

- Unified session control function:
  This function provides a method to control sessions in a unified signaling across multiple different mobile access networks. This function should also address connectionless services by providing light-weight signaling mechanism.
- Mobility control function:
  This function provides a unified mobility control across multiple different mobile access networks – same types of mobile access networks and heterogeneous mobile access networks.
- Multi-RAT control function:
  This function manages the resources from multiple wireless networks as an integrated virtual resource cooperating with multi-RAT coordination function in CGF.
- Policy management and QoS control function:
  This function controls the service policies and QoS control policies of user and each service flow based on the user and application identification.
- Identifier management function:
  This function manages all the identifiers of mobile devices including device identifier of 3GPP-types, MAC identifier of Wi-Fi, or other types of identifiers for user or device. All the identifiers need to be managed in such a way that the identifiers can be cross checked with each other.

9.3 IMT-2020 interworking scenarios with legacy IMT networks

The current IMT networks has been developed through an evolutionary path starting from circuit-based wireless networks to full IP-based mobile networks. The evolutionary approach of mobile networks with the aim of providing seamless continuity among the networks has significantly increased the complexity of the current mobile packet core.

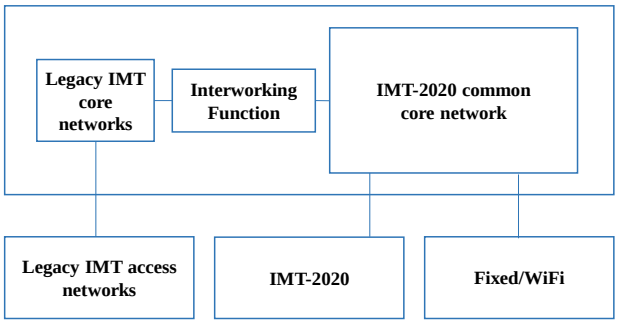
The IMT-2020 core network is envisioned as an access agnostic common core network which will bring a significant change to make it easier to interwork with a variety of new radio access technologies. However, the tight coupled interworking of an IMT-2020 network with legacy IMT networks will still bring the similar or even more complexity.

IMT-2020 network is expected to be designed with the consideration of Wi-Fi and fixed networks with emerging IMT-2020 RAN as basic access networks which will be seamlessly integrated into an

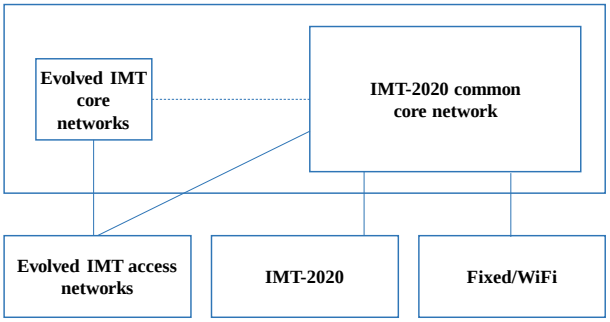
IMT-2020 network. However, we need to provide existing IMT networks a migration path to IMT-2020.

As the first option, legacy IMT networks are connected to an IMT-2020 network through an interworking function, which means that the IP session continuity between IMT-2020 and legacy networks can be provided but service continuity should be supported at higher layer solutions.

In the second option, evolved IMT networks are directly connected to an IMT-2020 network while providing backward compatibility for mobile devices that only support legacy IMT networks. In this case, IP session continuity as well as service continuity between IMT-2020 and legacy networks can be provided.



(a) Option 1



(b) Option 2

Figure 4. IMT-2020 interworking options.

10 Contributors (in Alphabetical Order)

This is the list of all contributors who submitted any written form of comments or contributions.

- Ghani ABBAS, Ericsson
- Seong Gon CHOI, Chungbuk University
- Prasan DE SILVA, Spark New Zealand
- Takash EGAWA, NEC Corporation
- Alex GALIS, University College London
- Chul-Soo KIM, Inje University
- Namseok KO, ETRI
- Kyunghee Daniel LEE, Pai Jai University
- Oscar LOPEZ-TORREZ

– Dharmendra MISRA, Cognizant
– Noik PARK, ETRI
– Sherry SHEN, Nokia Networks
– Prakash SUTHAR, Cisco Systems
– Toshitaka TSUDA, Waseda University
– Jian WANG, Ericsson

Acknowledgement

This work was partially supported by the European Union 7th Framework Program DOLFIN project ("Data centres optimization for energy-efficient and environmentally friendly internet"; http://www.dolfin-fp7.eu), the EU H2020 5G PPP projects: 5GEX ("5G Multi-Domain Exchange" ; https://5g-ppp.eu/5GEx) and SONATA ("Service Programing and Orchestration for Virtualized Software Networks"; https://5g-ppp.eu/sonata/) and the European Space Agency (ESA) INSTINCT project ("Scenarios for integration of satellite components in future networks"; https://artes.esa.int/projects/instinct).

Appendix II

Network Softwarization for IMT-2020 networks

Editor's Note: Appendix II was produced during the FG-IMT 2020 focus group in order to investigate gaps in standardization related to IMT-2020. While the request from SG-13 was to deliver a report outlining standardization gaps, the consensus of the focus group was that the working documents produced and used during the focus group work contained useful information for future work and should be captured. Note, however, the focus group concentrated on producing accurate descriptions of the standardization gaps in the main body of this document; some minor errors may exist in the appendices. They are, however, the output of the focus group but are provided for information only.

Editor's Note: This appendix uses clause references in a form usually associated for normative text. This is maintained for this report to align with references made in the main body of this report.

This document is the deliverable of Network Softwarization for IMT-2020 networks. It contains all the revision proposed in the editorial meeting of 27-30 October2015.

The structure of this document is as follows: Section 2 presents the relevant references; Section 3 presents the terms defined in this document and the terms defined elsewhere used in this document; Section 4 presents the abbreviations; Section 5 states that there are no conventions used in this document; Section 6 presents the network softwarization and its main features including: Motivations; Relevant Use Cases; Overview of network softwarization; Characteristics of network softwarization; Energy management aspects of Network softwarization and Economic incentive aspects of network softwarization. It includes the related standardisation gaps analysis; Sections 7, 8, 9 present the technology areas that would utilize network softwarization technologies in 5G mobile networks (e.g. Integrated Network Management and Orchestration; Mobile Edge Computing; Distributed Cloud for Service Providers; In-network Processing; Resource Usage Optimization; Resource Abstraction; Migration towards softwarized networks, RAN Virtualization; Capability Exposure) including the related standardization gaps analysis; Clause 17.1 presents satellite networks aspects of network softwarization; Clause 17.2 presents legacy reduction techniques; Clause 17.3 presents Standardization efforts of Mobile Edge Computing; Clause 17.4 presents supplemental figures on network softwarization; Acknowledgement and Contributors list are completing this document.

Draft deliverable of Network softwarization for IMT-2020 networks

1 Scope

This report explores the technology areas of study in network softwarization [6]for resolving challenges for realizing IMT-2020 visions in order to identify a gap to be filled by ITU-T Study Group 13 (SG13) studies. It covers the non-radio parts [7]of the IMT-2020 networks. In this document the terms 'IMT-2020' and '5G' are used interchangeably.

2 References

The following ITU-T Recommendations and other references contain provisions that, through reference in this text, constitute provisions of this Recommendation. At the time of publication, the editions indicated were valid. All Recommendations and other references are subject to revision; users of this Recommendation are therefore encouraged to investigate the possibility of applying the most recent edition of the Recommendations and other references listed below. A list of the currently valid ITU-T Recommendations is regularly published.

The reference to a document within this Recommendation does not give it, as a stand-alone document, the status of a Recommendation.

[ITU-T Y.3001] Recommendation (2012) - Future networks: Objectives and design goals - http://www.itu.int/rec/T-REC-Y.3001-201105-I;

[ITU-T Y.3021] Recommendation (2012) - Framework of Energy Saving in Future Networks - http://www.itu.int/rec/T-REC-Y.3021-201201-I;

[ITU-T Y.3031] Recommendation (2012) - Identification framework in future networks - http://www.itu.int/rec/T-REC-Y.3031-201205-I;

[ITU-T Y.3011] Recommendation ITU-T Y.3011 (2012), *Framework of network virtualization for future networks* – http://www.itu.int/rec/T-REC-Y.3011-201201-I;

[ITU-T Y.3012] Recommendation ITU-T Y.3012 (2014), *Requirements of network virtualization for future networks* - https://www.itu.int/rec/T-REC-Y.3012;

[ITU-T Y.3300] Recommendation (2014) - Framework of software-defined networking - https://www.itu.int/rec/T-REC-Y.3300-201406-I;

[ITU-T Y.3500] Recommendation (2014) - Cloud computing – Overview and vocabulary- http://www.itu.int/rec/T-REC-Y.3500-201408-I;

[ITU-T Y.3510] Recommendation (2013) - Cloud computing infrastructure requirements – https://www.itu.int/ITU-T/rec/T-REC-Y.3510;

[ITU-T Y.3502] Recommendation (2014) - Cloud computing - Reference architecture – https://www.itu.int/ITU-T/rec/T-REC-Y.3502;

---

[6] The definition of the terminology "Network Softwarization" is described in 3.2.

[7] Note that we include in the scope the technology areas that span across the boundary between wireless and wired networks, e.g., C-RAN (Cloud Radio Access Network), end-to-end slices over resources including RATs (Radio Access Technologies), etc.

[ITU-T Y.3511] Recommendation (2014) - Framework of inter-cloud computing - https://www.itu.int/rec/T-REC-Y.3511;

[ITU-T Y.3512] Recommendation (2014) - Cloud computing - Functional requirements of Network as a Service - http://www.itu.int/rec/T-REC-Y.3512-201408-P;

[ITU-T Y.3513] Recommendation (2014) - Cloud computing - Functional requirements of Infrastructure as a Service - http://www.itu.int/rec/T-REC-Y.3513-201408-I ;

[ETSI NFV] ETSI ISG NFV, Network Functions Virtualisation, http://portal.etsi.org/portal/server.pt/community/NFV

[IETF RFC 3746] IETF RFC 3746 (2004), Forwarding and Control Element Separation (ForCES) Framework.

[IETF RFC 7149] IETF RFC 7149 (2014), Software-Defined Networking: A Perspective from within a Service Provider Environment.

[IETF SFC] IETF Service Function Chaining (sfc) working group, http://datatracker.ietf.org/wg/sfc/charter/

[ONF] Open Networking Foundation, "OpenFlow/Software-Defined Networking (SDN)," https://www.opennetworking.org/.

[SDN-WS Nakao] "Deeply Programmable Network", Emerging Technologies for Network Virtualization, and Software Defined Network (SDN), ITU-T Workshop on Software Defined Networking (SDN), http://www.itu.int/en/ITU-T/Workshops-and-Seminars/sdn/201306/Documents/KS-Aki_Nako_rev1.pdf.

[Programmable Networks - Galis]–"Programmable Networks for IP Service Deployment" ISBN 1-58053-745-6, pp450, June 2004, Artech House Books, http://www.artechhouse.com/International/Books/Programmable-Networks-for-IP-Service-Deployment-1017.aspx,

[Cloud Survey –Heilig] - "A Scientometric Analysis of Cloud Computing Literature"- a review of approx. 25,000 papers] - IEEE Transactions on Cloud Computing, Volume: PP, Issue: 99, 30 April 2014, ISSN: 2168-7161; DOI: 10.1109/TCC.2014.2321168

[Draft ETSI GS MEC 002 V0.4.2(2015-07)] Mobile-Edge Computing (MEC); Technical Requirements

[ETSI NFV MANO] Network Functions Virtualisation (NFV) Management and Orchestration- http://www.etsi.org/deliver/etsi_gs/NFV-MAN/001_099/001/01.01.01_60/gs_nfv-man001v010101p.pdf

[FG IMT-2020 WS 5GMF Nakao] – Akihiro Nakao, "Overview of network softwarization and adoption to 5G", ITU-T Focus Group on IMT-2020, Pre-meeting workshop on network, softwarization, Turin, Italy, 21 September 2015

[FG IMT-2020 WS Galis] – Alex Galis, "Challenges in network softwarization", ITU-T Focus Group on IMT-2020, Pre-meeting workshop on network softwarization, Turin, Italy, 21 September 2015

[FG IMT-2020 WS Manzalini] - Antonio Manzalini, Telecom Italia / IEEE SDN Committee: "R&D status of network softwarization", ITU-T Focus Group on IMT-2020, Pre-meeting workshop on network softwarization, Turin, Italy, 21 September 2015

[FG IMT-2020 WS Wang] Yachen Wang, China Mobile: "Key technologies to support network softwarization", ITU-T Focus Group on IMT-2020, Pre-meeting workshop on network softwarization, Turin, Italy, 21 September 2015

[FG IMT-2020 WS Tsuda] Toshitaka Tsuda, Waseda University: “CCN implementation by network softwarization”, ITU-T Focus Group on IMT-2020, Pre-meeting workshop on network softwarization, Turin, Italy, 21 September 2015

[ITU-R IMT Vision] ”Framework and overall objectives of the future development of IMT for 2020 and beyond,” ITU-R、Document 5/199-E、19 June 2015.

3 Terms defined in this report

3.1 Terms defined elsewhere

This Recommendation uses the following terms defined elsewhere:

3.1.1 future network (FN) [ITU-T Y.3001]: A network able to provide services, capabilities, and facilities difficult to provide using existing network technologies. A future network is either:

a) A new component network or an enhanced version of an existing one, or

b) A heterogeneous collection of new component networks or of new and existing component networks that is operated as a single network.

NOTE – The plural form "Future Networks" (FNs) is used to show that there may be more than one network that fits the definition of a future network.

3.1.2 network virtualization [ITU-T Y.3011]: A technology that enables the creation of logically isolated network partitions over shared physical networks so that heterogeneous collection of multiple virtual networks can simultaneously coexist over the shared networks. This includes the aggregation of multiple resources in a provider and appearing as a single resource

3.1.3 software-defined networking [ITU-T Y.3030]: A set of techniques that enables to directly program, orchestrate, control and manage network resources, which facilitates the design, delivery and operation of network services in a dynamic and scalable manner.

3.1.4 energy saving within networks [ITU-T Y.3021]: This is where network capabilities and their operations are set up in a way that allows the total energy for network equipment to be systematically used in an efficient manner, resulting in reduced energy consumption, compared with networks that lack these capabilities and operations.

NOTE – Network equipment includes routers, switches, equipment at the terminating point e.g., optical network units (ONUs), home gateways, and network servers such as load balancers and firewalls. Network equipment is typically composed of various components such as switching fabric, line cards, power supply, and cooling.

3.1.5 cloud service customer [ITU-T Y.3501]: A person or organization that consumes delivered cloud services within a contract with a cloud service provider.

3.1.6 cloud service provider [ITU-T Y.3501]: An organization that provides and maintains delivered cloud services.

3.1.7 management system [ITU-T M.60]: A system with the capability and authority to exercise control over and/or collect management information from another system.

3.1.8 device [ITU-T Y.3021]: This is the material element or assembly of such elements intended to perform a required function.

3.1.9 equipment ITU-T Y.3021]:A set of devices assembled together to form a physical entity to perform a specific task.

3.1.10 virtual resource [ITU-T Y.3011]: An abstraction of physical or logical resource, which may have different characteristics from the physical or logical resource and whose capability may not be bound to the capability of the physical or logical resource.

3.1.11 logical resource [ITU-T Y.3011]: An independently manageable partition of a physical resource, which inherits the same characteristics as the physical resource and whose capability is

bound to the capability of the physical resource.

NOTE – "independently" means mutual exclusiveness among multiple partitions at the same level.

3.1.12 resource management [ITU-T Y.3520]:: The most efficient and effective way to access, control, manage, deploy, schedule and bind resources when they are provided by service providers and requested by customers.

3.1.13 hypervisor [ITU-T Y.3510]: A type of system software that allows multiple operating systems to share a single hardware host.

NOTE –Each operating system appears to have the host's processor, memory and other resources, all to itself

3.1.14 virtual machine [DMTF OVF]: The complete environment that supports the execution of guest software.

3.1.15 identifier [ITU-T Y.2091]: An identifier is a series of digits, characters and symbols or any other form of data used to identify subscriber(s), user(s), network element(s), function(s), network entity(ies) providing services/applications, or other entities (e.g., physical or logical objects).

3.1.16 locator (LOC) [ITU-T Y.2015]: A locator is the network layer topological name for an interface or a set of interfaces. LOCs are carried in the IP address fields as packets traverse the network.

NOTE – In this Recommendation, locators are also referred to as location IDs.

3.1.17 node ID [ITU-T Y.2015]: A node ID is an identifier used at the transport and higher layers to identify the node as well as the endpoint of a communication session. A node ID is independent of the node location as well as the network to which the node is attached so that the node ID is not required to change even when the node changes its network connectivity by physically moving or simply activating another interface. The node IDs should be used at the transport and higher layers for replacing the conventional use of IP addresses at these layers. A node may have more than one node ID in use.

NOTE – This Recommendation specifies a node ID structure.

3.1.18 name [ITU-T Y.2091]: A name is the identifier of an entity (e.g., subscriber, network element) that may be resolved/translated into address.

3.1.19 domain [ETSI NFV MANO]:

Administrative domain is a collection of systems and networks operated by a single organization or administrative authority. Infrastructure domain is an administrative domain that provides virtualised infrastructure resources such as compute, network, and storage or a composition of those resources via a service abstraction to another Administrative Domain, and is responsible for the management and orchestration of those resources.

NOTE1: Different networks and different parts of a network may be built as different domains using separate technologies or having different control paradigms,

NOTE2: Different networks and different parts of a network may be owned by a single administration creating an *administrative domain*. Services are enabled and managed *over multiple administrations or over multi-domain single administration.*

NOTE3: *Multitenancy domain* refers to set of physical and /or virtual resources in which a single instance of a software runs on a server and serves multiple tenants. A tenant is a group of users who share a common access with specific privileges to the software instance. A service or an application may be designed to provide every tenant a dedicated share of the instance including its data, configuration, user management, tenant individual functionality and non-functional properties.

3.2 Terms defined in this recommendation

This Recommendation defines the following terms:

3.2.1 Network Softwarization: An overall transformation trend for designing, implementing, deploying, managing and maintaining network equipment and/or network components by software programming, exploiting the natures of software such as flexibility and rapidity all along the lifecycle of network equipment / components, for the sake of creating conditions enabling the re-design of network and services architectures, optimizing costs and processes, enabling self-management and bringing added values in network infrastructures.

4 Abbreviations and acronyms

This Recommendation uses the following abbreviations and acronyms:

SDN Software Defined Networking

NFV Network Virtualization

LINP a logically isolated network partitions

5 Conventions

6 Network softwarization

6.1 Motivations

In preparing for wired networks in coming IMT-2020 era, a variety of changes in information and telecommunication environment and social requirements related to network infrastructure should be taken into account. Typical examples of these are as follow.

1. Video traffic already dominates the mobile communication traffic and still increasing. This requires the network architecture having efficient video on demand delivery mechanism in terms of network/server congestion reduction and shorter response time.

2. In many countries, disaster resilience is a key concern. Since network systems are a life line infrastructure, they are expected to be robust. Increased network flexibility helps to respond this request.

3. IoT and big data processing are booming. Future networks should provide with functions that fit to efficient big data processing systems.

4. One of the major requirements for 5G is the very short delay. Total design including data processing and service provisioning is necessary to fulfil the requirement.

5. SDN/NFV concept is expanding as a possible key technology for future networks.

To satisfy these requirements, it is appropriate that the 5G networks be built upon a new network architectural model which provides flexibility in terms of network topology and functions, enabling the creation of multiple logical networks dedicated to the support of specific services, the introduction of emerging architectural paradigms such as ICN/CCN, and the realization of in-network data processing and service provisioning provide by network nodes.

This flexibility objective can be achieved via Network softwarization. With the adoption of SDN/NFV, software programmability of network nodes will expand, which in turn makes it feasible to run information processing and service software on network nodes.

6.2 Relevant use cases

6.2.1 Support of various applications

In the 5G era, many new application services are expected to emerge to satisfy diverse needs and requirements of users by leveraging innovative multimedia applications and telecommunication technologies.

Figure 1 indicates that, from users' perspectives, required capabilities vary depending on applications. Four typical applications are used to illustrate the different capabilities required by respective application.

(i) Video streaming
(ii) Virtual reality
(iii) M2M communication (i.e. sensors)
(iv) Autonomous driving (i.e. collision avoidance)

As users' requirements differ depending on applications, the network does not necessarily have to exhibit maximum performance capabilities, instead, the network resources should be used efficiently depending on applications.

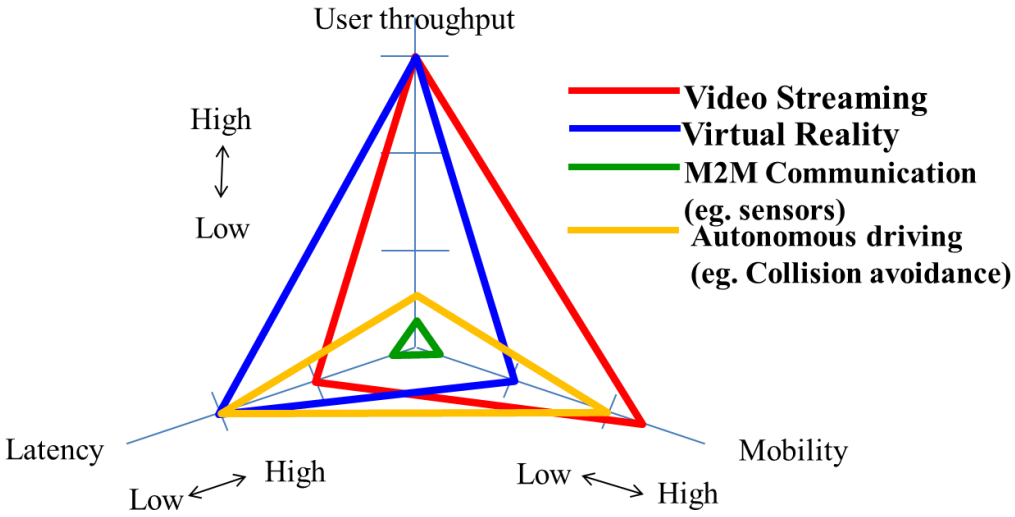


Figure 1 Diversified required capabilities from users' perspective depending on applications

Gap analysis

It is envisioned that such an infrastructure that efficiently supports a diversified set of application requirements across end-to-end paths, ranging from M2M communication, to autonomous and collaborative driving, virtual reality and video streaming, etc. Network softwarization technologies including SDN, NFV and their extensions for supporting 5G mobile networks are expected to provide slicing capability both in wired and wireless parts of communication infrastructure, so that each slice provides an isolated environment to efficiently accommodate individual applications meeting specific requirements. The slice should be capable of dynamically adjusting resources to meet the application requirements. The network infrastructure is expected to provide extreme flexibility to support those different capabilities with reasonable cost.

6.2.2 Use of ICN for Inter-function transport

In the future 5G network, where a high degree of network function virtualization is expected either explicitly in NFV or in network slices, the use of an ICN protocol for inter-function communications is an attractive option because it decouples the location of services from the service request. The concept of using ICN in NFV and SDN is studied in several academic works[8][9][10][11]. ICN is well suited for service oriented routing, because each element in a service chain can name the next element in the service chain without the need for an external name resolver or manual configuration.

The initial deployment of virtualized functions inside a network Slice could use ICN technologies immediately, as these are green-field services realized within a provider network [12]. Initial implementations may require running ICN as a transport protocol over IP due to initial lack of hardware support for native ICN transport, but this would be a transitional situation. Even when running over an IP network layer, ICN services can still provide robust communications and endpoint discovery using common existing IP techniques (e.g. mDNS, DNS SRV records, and distribute rendezvous techniques, among others).

As data-plane programmable equipment enters the marketplace, native ICN slices and transport can offer high-performance ICN/CCNx services. For example, abstracted 5G services in a CCNx slice, such as MME, S-GW, and P-GW, could be implemented in software, in software with hardware assist, or in deep-programmable hardware. In each case, the abstracted Slice service looks the same, though each would offer different performance curves in terms of latency, throughput, and capacity. Likewise, for inter-NFV communications, native wire-speed equipment, which is being demonstrated in 2015 by some manufacturers, would improve throughput and flexibility compared to using an IP-overlay, but should not limit such deployments within the 5G timeframe.

The advantage of using ICN as the inter-function transport protocol, even in a transitional period over IP, is that it positions carriers to move to an ICN native approach, which can discovery, recover, and utilize carrier networks efficiently without centralized bottlenecks or points of failure. ICN approaches also position the carrier for dynamic service mobility and deployment without needing to track an IP underlay.

Gap analysis

There are both academic and commercial research activities for defining emerging network architectures that do not assume the underlay network runs TCP/IP protocols. A representative example of such network architectures is ICN (Information Centric Networking). Although the

---

[8] Latre, Steven, et al. "The fluid internet: service-centric management of a virtualized future internet." *Communications Magazine, IEEE* 52.1 (2014): 140-148.

[9] Ravindran, Ravishankar, et al. "Towards software defined ICN based edge-cloud services." *Cloud Networking (CloudNet), 2013 IEEE 2nd International Conference on*. IEEE, 2013.

[10] TalebiFard, Peyman, et al. "Towards a context adaptive ICN-based service centric framework." *Heterogeneous Networking for Quality, Reliability, Security and Robustness (QShine), 2014 10th International Conference on*. IEEE, 2014.

[11] Nakao, Akihiro. "Software-defined data plane enhancing SDN and NFV." *IEICE Transactions on Communications* 98.1 (2015): 12-19.

[12] Nakao, Akihiro. "Application Specific Slicing for MVNO through Software-Defined Data Plane Enhancing SDN." *IEICE Transactions on Communication, Vol. E98-B, No. 11, 2015.*

current state of ICN could run on top of TCP/IP as an overlay, inherent benefits of the architecture can be achieved if implemented natively, i.e., directly on top of the underlay, e.g., L1 or L2 networks. There exists a gap in supporting such emerging network architecture in the current network technology, especially when it uses such an emerging network architecture in the context of heterogeneous service delivery and function chaining. Network softwarization provides slicing such that such multiple emerging network architectures could be realized within individual slices.

Gap analysis

1. To fully realize benefit, need native ICN routing within a slice (e.g., on CRAN and backhaul). There could be a case where inter-slice ICN routing is useful.
2. NFV and Slice implementations would need to use ICN technologies in their communications models.

6.2.3 Use of ICN function state migration

ICN is a good technology choice for function state migration and execution context migration in future 5G networks[13][14][15]. Content Centric Networking (CCNx) facilitates process migration while enabling many desirable features such as strong checkpointing and data de-duplication. Not all migration techniques require strong checkpointing, and in those cases CCNx offers a faster and weaker naming technique that allows pages or blocks to go dirty in a checkpoint.

CCNx offers an intuitive naming of resources that are part of a service or execution context migration and building checkpoints around those resources for consistent state transfer.

De-duplication is a technique where only one copy of data exists and it is shared between multiple instances. CCNx allows resources to be de-duplicated both *within* and *between* virtual machine instances. For example, in the previous discussion about using hash names for resources, if two disk blocks, for example, have the same hash value they will refer to the same Content Object. Only the block index in the CCNx manifest will be different.

A VM hypervisor may also share blocks between VMs. When generating the names used to fetch a checkpoint, the source migration agent running in the source hypervisor could use a name like {/nyc/host7, hash = 0x63223…} so any instance or any component can share the same data. Assume that the memory page size and the disk block size are the same. Then that name for hash 0x63223… could be both a disk block and a RAM page of the same data (e.g. a shared library code section). Because the manifest can point to different name prefixes for each hash and can indicate the virtual resource of that hash, we can have the same physical bytes used for many purposes. This approach may also be applied when page and block sizes are not the same by using smaller units of naming.

---

[13] Taleb, Tarik, and Adlen Ksentini. "Follow me cloud: interworking federated clouds and distributed mobile networks." *Network, IEEE* 27.5 (2013): 12-19.

[14] Karimzadeh, Morteza, Triadimas Satria, and Georgios Karagiannis. "Utilizing ICN/CCN for service and VM migration support in virtualized LTE systems." (2014): 84.

[15] Crowcroft, Jon. "SCANDEX: Service Centric Networking for Challenged Decentralised Networks."

Gap analysis

1. An ICN (CCNx) transport within the CRAN or carrier network to facilitate ICN approaches to state migration. The ICN technology could operate over an IP network during a transitional period.
2. NFV or Slice component implementations would need to use ICN as their transport technology to benefit from the name-based design and transfer features.

6.2.4 ICN removes host endpoint abstraction from application data

In a traditional IP-based network, the Internet Protocol adds a level of abstraction to communications endpoints by assigning them a location-dependent name. When applications wish to communicate, they must rendezvous those host addresses – such as with DNS or SIP or some other well-known means – before communication can take place. Because the rendezvous is done outside the network layer, the rendezvous protocol must employ its own means to determine locality – such as with ping triangulation – to determine which replica to use. Some applications use IP anycast addressing to move rendezvous back in to the network layer and realize those benefits. CCNx, as an ICN protocol, naturally keeps rendezvous on addresses within the network layer so all applications can benefit from localized services without needing to add on additional rendezvous layers with their own localization protocols.

ICN may be used as a de-abstraction layer for virtualized functions: using direct function naming in ICN means the network can move functions and change routing without needing to update intermediate abstractions of endpoint identification. For example, a single host IP address might hide many virtualized functions, so it may not be possible to directly move an IP address. One would need orchestration to inform components of a new socket endpoint, which could result in service interruption during the time when a function has finished migration and the time when an existing prior service is notified of a new service endpoint. With CCNx, the orchestration does not need to inform prior components of a new service address, it only needs to update the named routing to the new location.

Because CCNx, as an example ICN protocol, is not tied to the P-GW identity – such as for the source endpoint address – it means that CCNx is well suited for multiple P-GW egress. Service frameworks, such as Mobile Edge Computing, could realize significant simplification by using a CCNx approach for multiple P-GW egress without needing to assign the UE multiple identities or using layers of address translation.

Gap analysis

1. To fully realize benefit, need native ICN routing within the CRAN and backhaul. ICN routing within a slice is insufficient.
2. Deep data plan programmability, extended from SDN and NFV, and Slice implementations would need to use ICN technologies in their communications models.
3. The CRAN would need to offer ICN routing, such as from a SIPTO P-GW exit at the eNodeB. It is possible to run ICN as an IP overlay, but a native ICN routing would be better.

6.3 Overview of network softwarization

Network Softwarization is an overall transformation trend for designing, implementing, deploying, managing and maintaining network equipment and/or network components by software programming, exploiting the natures of software such as flexibility and rapidity all along the lifecycle of network equipment / components, for the sake of creating conditions enabling the re-design of network and services architectures, optimizing costs and processes, enabling self-management and bringing added values in network infrastructures.

The terminology, Network Softwarization, was first introduced in Academia, NetSoft 2015, the first IEEE Conference on Network Softwarization, to include broader interests regarding Software Defined Networking (SDN) and Network Functions Virtualisation (NFV), Network Virtualization, Mobile Edge Computing, Cloud and IoT technologies.

6.4 Characteristics of network softwarization

Additional benefits from Network Softwarization include but are not limited to the following:

Network Softwarization enables global system qualities (e.g. execution qualities, such as usability, modifiability, effectiveness, security and efficiency; evolution qualities, such as testability, maintainability, reusability, extensibility, portability and scalability). Viable architectures for network softwarization must be carefully engineered to achieve suitable trade-offs between flexibility, performance, security, safety and manageability.

Network Softwarization provides a set of software techniques, methods and processes applicable to a heterogeneous assembly of component networks or an enhanced version of an existing grouping of network components that is operated as a single network.

Network Softwarization provides abstractions and programmability that are utilized to deliver extreme flexibility in networks to support variety of applications and services, to accelerate service deployment and to facilitate infrastructure self-management.

Network Softwarization enables the following high-level characteristics and capabilities, their extensions, their trade-offs, unification and integration:

- *Network Virtualization (NV),* which enables virtualization of network resources,
- *Network Function Virtualization (NFV),* which permits virtualization of software-based network functions. Instead of installing and managing dedicated hardware devices for these networking and servicing functions, they are instead realized as software components and deployed on commodity or special hardware infrastructures.
- *Network Programmability* empowers the fast, flexible, and dynamic deployment of new network and management services executed as groups of virtual machines in the data plane, control plane, management plane as service plane in all segments of the network. Programmable networks are networks that allow the functionality of some of their network elements to be programmable dynamically. These networks aim to provide easy introduction of new network services by adding dynamic programmability to network devices such as routers, switches, and applications servers. Dynamic programming refers to executable code that is injected into the execution environments of network elements in order to create the new functionality at run time. The basic approach is to enable trusted third parties (end users, operators, and service providers) to inject application-specific services (in the form of code) into the network. Applications may utilize this network support in terms of optimized network resources and, as such, they are becoming network aware. As such the behaviour of network resources can be customized and changed through a standardized programming interface for network control, management and servicing functionality. The key question is: how to exploit

this potential flexibility for the benefit of both the operator and the end user without jeopardizing the integrity of the network. The answer lies in the promising potential that emerges with the advent of programmable networks in the following aspects: Rapid deployment of large number of new services and applications; Customization of existing service features; Scalability and operational cost reduction in network and service management; Independence of network equipment manufacturer; Information network and service integration; Diversification of services and business opportunities.

- *Software-Defined Networking* (*SDN),* which allows network control to be separated from the forwarding plane and allows for a flexible management of the network resources which facilitates the design, delivery and operation of network services in a dynamic and scalable manner.
- *Software-defined Network Clouds - Cloudification of networking and servicing functions*, which enables ubiquitous network access to a shared services and shared pool of configurable computing, connectivity and storage resources. It provide users and providers with various capabilities to process and store their data and services in data centers. It relies on sharing of resources to achieve coherence and economies of scale, similar to a utility (like the electricity grid) over a network. It uses virtualization concepts such as abstraction, pooling, and automation to all of the connectivity, compute and storage to achieve network services. It could take also the kind of mobile edge computing architecture where cloud-computing capabilities and an IT service environment are available at the edge of the mobile network or fog architecture that uses one or a collaborative multitude of end-user clients or near-user edge devices to execute a substantial amount of services (rather than in cloud data centers), communication (rather than routed over the internet backbone), and control, configuration, measurement and management.

6.5 Requirements for 5G specific network

(1) *Harmonization of SDN and NFV* - Coordination of the current SDN and NFV technologies for realizing 5G mobile network is required.

(2) 5G Extensions to the current SDN and NFV - 5G network needs extreme flexibility in supporting various applications and services with largely deferent requirements. Therefore, 5G specific extensions to the current SDN and NFV, especially pursuing even further and deeper agile software programmability is required. For example, SDN data plane could be enhanced to support deep programmability and NFV MEC needs light-weight management for extreme edge network functions, especially in the area of access network and user equipment (UE).

(3) Considerations for applicability of softwarization - Considering the trade-off between programmability and performance is required. Especially in 5G context, it is important to respect the performance improvement in wireless technologies. Therefore, it is necessary to clearly define the area and criteria for applicability of softwarization in the infrastructure.

(4) *Application driven 5G network softwarization* - 5G mobile network is indispensable communication infrastructure for various applications and services such as IoT/M2M and content delivery. Rapid emergence of applications and services enabled in 5G mobile network must be considered in designing and developing the infrastructure.

(5) *5G network softwarization energy characteristics* - The architecture design, resulting implementation and operation of 5G network softwarization are recommended to minimize their environmental impact, such as explicit energy closed control loops that optimizes energy consumptions and stabilization of the local smart grids at the smart city level.

(6) *5G network softwarization management characteristics* - The architecture design, resulting implementation and operation of 5G network softwarization are recommended to included uniform and light-weight in-network self-organization, deeper autonomy, and autonomicity as its basic enabling concepts and abstractions applicable to all components of 5Gnetwork.

(7) *5G network softwarization economic characteristics* - The architecture design, resulting implementation and operation of 5G network softwarization are recommended to consider social and economic issues to reduce as target 50% the \systems and subsystems lifecycle and operational costs in order for them to be deployable and sustainable, to facilitate appropriate return for all actors involved in the networking and servicing ecosystem and to reduce their barriers to entry.

6.6 Network softwarization in 5G mobile networks

6.6.1 Network softwarization

In 5G or IMT-2020 networks, the terminology Network Softwarization is used with the intention to introduce various requirements of programmable software defined infrastructure, especially specific extension for 5G mobile networks.

The basic concept of the Network Softwarization is “Slicing” as defined in [ITU-T Y.3011], [ITU-T Y.3012]. Slicing allows logically isolated network partitions (LINP) with a slice being considered as a unit of programmable resources such as network, computation and storage. Considering the wide variety of application domains to be supported by 5G or IMT-2020 network, it is necessary to extend the concept of slicing to cover a wider range of use cases than those targeted by the current SDN/NFV technologies, and the need to address a number of issues on how to utilize slices created on top of programmable software defined infrastructure.

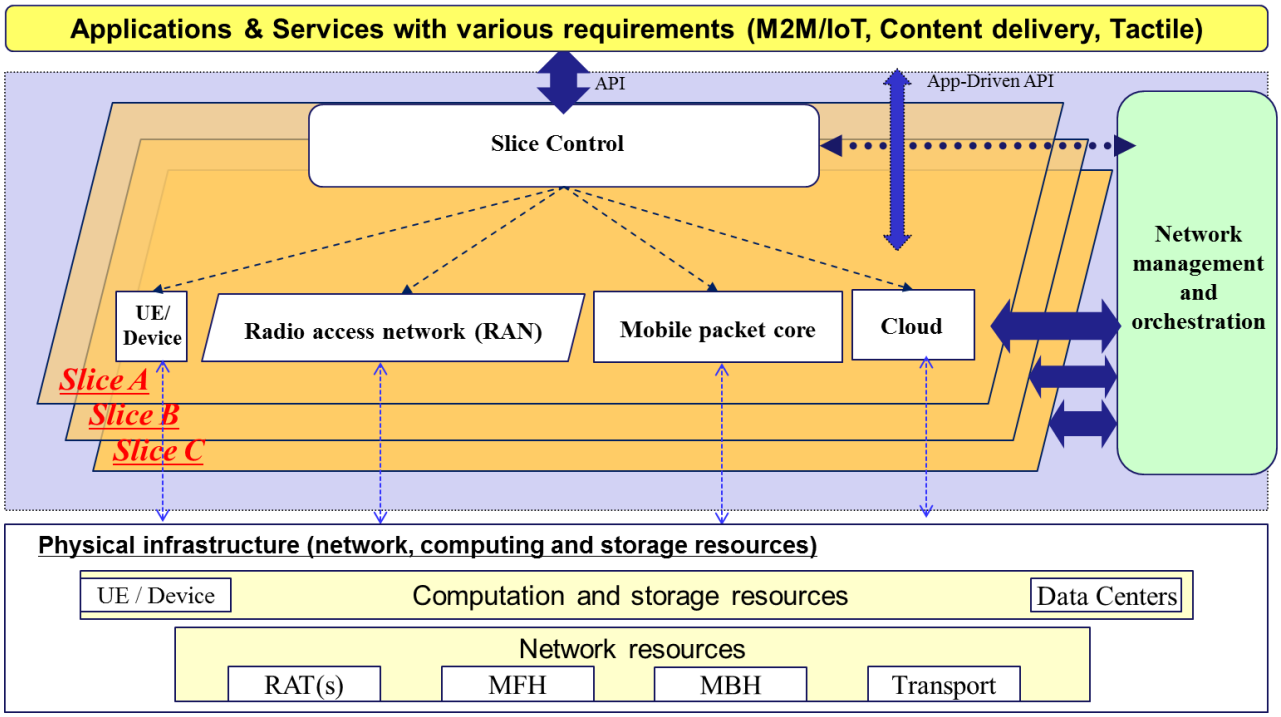


Figure 2 Network softwarization view of the 5G mobile networks

Figure 2 illustrates the network softwarization view of 5G mobile networks, which consists of a couple of slices created on a physical infrastructure by the "network management and orchestration". A slice is the collection of virtual or physical network functions connected by links to create an end-to-end networked system. In this figure, the slice A consists of radio access network (RAN), mobile packet core, UE(User Equipment)/device and cloud, each of which are collection of virtual or physical network functions. Note that the entities are shown rather symbolically and links are not described in Figure 2 for simplicity. The "network management and orchestration" manages the life cycle of slices: creation, update and deletion. It also manages the physical infrastructure and virtual resources, which are abstraction of physical resources. The physical infrastructure consists of computation and storage resources that include UE/devices (e.g. sensors) and data centers, and network resources that include RATs, MFH, MBH and Transport. It should be noted that both computation/storage resources and network resources are distributed and are available for creating virtual network functions.

In addition, the virtualized network functions and network programmability functions assigned to a slice are controlled by the "slice control". It oversees the overall end-to-end networked system by configuring its entities appropriately. It may include network layer control, and service/application layer control in some cases making 5G network control being service-aware. It depends on the requirements presented for the end-to-end networked system as exemplify by the envisaged natively supported ICN facilities.

The orchestration is central to the control and it defined as the sequencing of management operations. A customer may send a request to the "network management and orchestration" with their own requirement of an end-to-end service and steps follows. As such the new 5G control is governing, synchronizing and enforcing the configuration of the natively supported NV / Slicing/ NFV and network programmability facilities in the fronthaul, backhaul, core networks, software-defined clouds and mobile edge computing. This involves:

- Support for on demand composition of network functions and capabilities
- Enforce required capability/capacity/security/elasticity/ adaptability/ flexibility "where and when needed"

Services are executed in one (or more) Slices (i.e. a slice is made of a set of VMs)

Step 1: Creating a slice

Based on the requirement presented, the "network management and orchestration" creates virtual or physical network functions and connects them as appropriate and instantiate all the network functions assigned to the slice.

Step 2: (Re-) Configuring the slice

The slice control take over the control of all the virtualized network functions and network programmability functions assigned to the slice, and (re-)configure them as appropriate to provide the end-to-end service.

6.6.2 Horizontal extension of slicing

6.6.2.1 End-to-end slicing

In 5G mobile networks, the end-to-end communication quality is an important requirement. Especially when wireless technologies are expected to advance, fixed network must support facilitating the advancement of wireless part of end-to-end communications. Therefore, it is natural to consider extending the slicing concept to end-to-end, i.e., from UE to Cloud. Issues in extending

slices in 5G mobile networks have then to be addressed, not only software defined infrastructure in a limited part of a network, but also the entire end-to-end path including UE and Cloud.

Gap analysis

The scope of the current SDN technology primarily focuses on the portions of the network such as data-centres, mobile and core networks. In 5G mobile network, it is necessary to consider end-to-end application quality and enablement through network softwarization platform. Therefore, there exists a gap between the current projection of SDN and NFV technology development and the requirements for end-to-end application quality. The infrastructure for 5G mobile networks is desired to support end-to-end control and management of slices and the composition of multiple slices, especially with consideration of slicing over RATs and fixed parts of end-to-end paths. This gap has been analyzed against what is defined in [ITU-T Y.3300].

6.6.2.2 End-to-end latency breakdown and programmability consideration

As presented at the pre-meeting of FG IMT-2020 on network softwarization at Turin, Italy, 21 September 2015 [FG IMT-2020 WS 5GMF Nakao], the Figures 3 show the breakdown of the end-to-end latency in the current mobile network architecture. These figures imply that the 5G mobile network architecture must be able to execute network functions and services at any part along the end-to-end communication in order to make the most of wireless latency reduction (targeted from 10msec to 1msec) described in ITU-R IMT Vision [ITU-R IMT Vision].

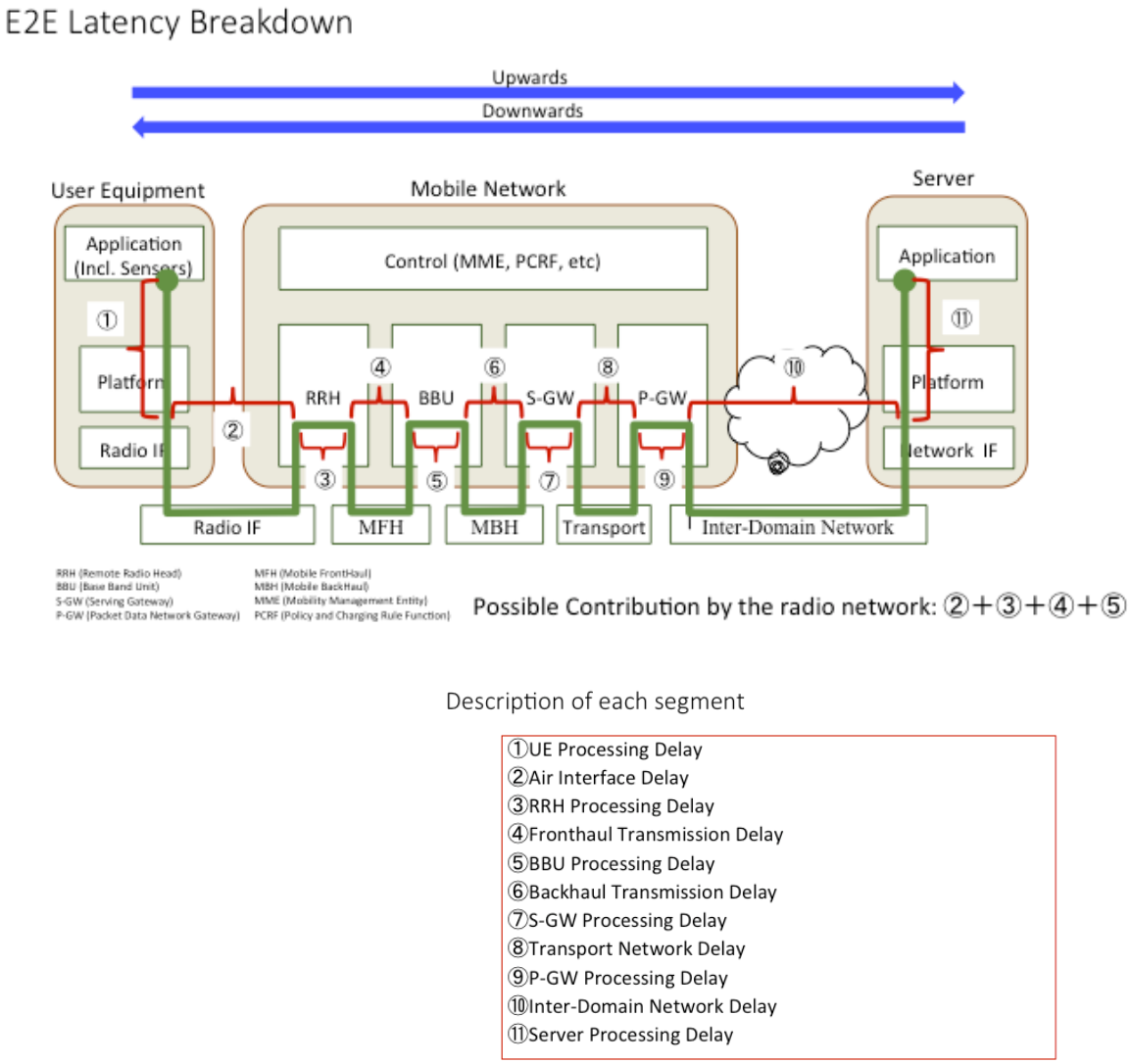


Figure 3 End-to-End Latency Breakdown of an example mobile (LTE) network

For 3GUMTS/LTE network 3GPP had carried out latency studies which is documented in specifications TR 25.912, TR25.913, TR36.912, TR36.913. 3GPP has carried out study for 5G network requirements, which is documented in TR22.891. Service providers are building LTE network to meet latency budget provided in 3GPP specification, so that services perform optimally. Latency studies carried out on many LTE deployed network demonstrate that 3GPP specifications provides adequate guidelines, however actual LTE network performance varies due to many variables and adjacent ecosystem.

For IMT-2020 network our recommendation is to carry out extensive latency study and provide guidelines for latency, packet loss and jitter etc. so that it provide optimal performance for many diverse applications. In order to structure the latency study framework, we suggest break-down network latency into three segments

A. Radio latency between User Equipment (UE) and base station (BS)

B. Network user-plane latency between base station, cloud-RAN, front/backhaul and mobile gateway handling user traffic. For D2X communications user-plane latency depend upon delay in base station

C. Network control-plane latency between base station, cloud-RAN and mobile gateways e.g. MME handling mobility management and other control function

6.6.3 Vertical extension of slicing (Data plane enhancement)

6.6.3.1 Deep data plane programmability

In 5G mobile networks, we must support various communication protocols (such as ones being invented) to support services such as Internet of Things (IoT) and for content delivery such as information centric networking (ICN) and content centric networking (CCN). Advanced infrastructure should provide capability of data-plane programmability and programming interfaces. Although SDN community only recently started tackling the issue of data plane programmability, in 5G mobile networks, it is significant to consider the vertical extension of SDN to support data plane programmability and programming interfaces and also of NFV to the very edge of the network closer to UE for emerging services and applications supported by new protocols, especially for IoT services and content delivery.

Gap analysis:

The current SDN technology primarily focuses on the programmability of the control plane, and only recently the extension of programmability to the data plane is being discussed in the research community and in ITU-T SG13 without well-defined use cases. For 5G mobile networking, there are several use cases for driving invention and introduction of new protocols and architectures especially at the edge of the network. For instance, the need for redundancy elimination and low latency access to contents in content distribution drives ICN at mobile backhaul networks.

Protocol agnostic forwarding methods such as Protocol Oblivious Forwarding (POF) discuss the extension to SDN addressing forwarding with new protocols. In addition, protocols requiring a large cache storage such as ICN needs new enhancement.

A few academic research projects such as P4[16] and FLARE[17] discuss the possibility of deeply programmable data plane that could implement new protocols such as ICN, but there is no standardization activity to cover such new protocols to sufficient extent.

Therefore, there exists a gap between the current projection of SDN and NFV technology development and the requirements for deep data plane programmability. The infrastructure for 5G mobile networks is desired to support deeper data plane programmability for defining new protocols and mechanisms. This gap has been analyzed against what is defined in [ITU-T Y.3300].

6.6.4 Considerations for applicability of softwarization

In 5G mobile networks, not every component of infrastructure may be defined by software and made programmable, considering the trade-off between programmability and performance. Therefore, according to the applications and services to be enabled, it is necessary to clearly define the role of

---

[16] Pat Bosshart, Dan Daly, Martin Izzard, Nick McKeown, Jennifer Rexford, Cole Schlesinger, Dan Talayco, Amin Vahdat, George Varghese, David Walker, “Programming Protocol-Independent Packet Processors”, http://arxiv.org/abs/1312.1719

[17] Nakao, Akihiro. "Software-defined data plane enhancing SDN and NFV." *IEICE Transactions on Communications* 98.1 (2015): 12-19.

hardware and software according to the possible application use cases when we softwarize the infrastructure.

Gap analysis:

SDN and NFV are primarily motivated by OPEX and CAPEX reduction and flexible and logically centralized control of network operations, and these technologies aim to focus on softwarization of everything everywhere possible to meet various network management and service objectives. Also the traffic classification is often per flow basis.

In 5G mobile networks, some applications have stringent performance requirements such as ultra-low latency and high peak rate while others may not require cost-effective solutions. A range of solutions exists from application driven software-based solutions executed on virtualization platform with hypervisor, container or bare metals, to complete hardware-assisted solutions. The former may need performance enhancement enabled by hardware-assisted solutions, while the latter may be facilitated by software-based solutions.

The infrastructure for 5G mobile network must support traffic classification performed not only by flow-basis but also by other metrics and bundles such as per-device and per-application basis so that we may apply software /hardware based solutions appropriately for individual use cases. Therefore, there exists a gap between the current projection of SDN and NFV technology development and the requirements for applicability of softwarization. This gap has been analyzed against what is defined in [ITU-T Y.3300].

6.6.5 End-to-end reference model for scalable operation

The softwarized networking systems should have sufficient levels of scalability in various aspects of functions, capabilities, and components. Firstly, the target range of number of instances should be considered, e.g. the service slices to be configured and to be in operation concurrently. The number of clients and service providers accommodated by each service slice is also an important measure for the practical deployment of the softwarized networking systems.

The main constraints for scalability of softwarized systems would be the dynamic behaviour of each slice and the control granularity of physical resources. The communication session established by mobile core, however, would be beyond the scope of this activity, because it requires a dedicated system for such an extraordinary multiple-state and real-time control, especially for the mobility handling. The coordination and isolation between these systems should be clearly defined.

Nevertheless, scalability for other types of sessions would be the issues of architectural modelling, including application services, system operation, or advanced network services.

In addition to the dimensions and dynamics of the softwarized systems, investigations would be required from the viewpoints of resiliency and inter-system coordination.

For resiliency, some new aspects might be considered other than traditional MTBF type faulty conditions. In case of disaster, for example, the fault localization, analysis, and recovery could be more complicated for such virtual systems with network softwarization. Miss-behaviors caused by human factors are also difficult to cope with the traditional operation architecture, because of the indirectness of virtual system operation.

The inter-system coordination architecture should be clearly structured and modelled for efficient standardization and for scalability evaluation of the softwarized networking systems. There might be two categories of the coordination, namely horizontal and vertical. The horizontal coordination is for between slice, cloud systems, and UE, the end-to-end system coordination in other word. Two types of vertical coordination could be distinguished. One is for slice and service provider through APIs

and another is for virtual and physical resource coordination aimed to efficient resource handling through policy and analytics.

In summary, the softwarized network systems should have sufficient levels of scalability in the components;

- The number of instances/service slices to be supported
- Series of capabilities provided by service slices
- The number of service sessions to be handled concurrently
- Dynamic behaviour of the instances and slices
- Granularity of resource management, especially for policy control and/or analytics
- Resiliency for various faulty conditions
- Intra-slice coordination among the end-to-end resources
- Inter-slice coordination, specifically with various external systems.

Gap analysis

Intensive studies are required on both the dimension and the dynamic behavior of softwarized networks, since such highly virtualized systems will have an enormous number of instances and reactions are not easy to extrapolate from the current physical systems.

The virtualized resource handling must be the essential part of the scalable and novel operation architecture, which potentially improves conventional network operations and possibly even up to the level of supporting disaster recovery by using softwarized network resiliency and recovery of/with the virtualized systems both in a single domain and in multiple domains.

One of the benefits of 5G systems should be the end-to-end QoE management, however, this capability will be established on the complex interaction between the virtualized systems including UEs, Cloud Systems, Applications, and the softwarized network systems. The softwarized network system itself will be composed of various virtualized subsystems. An appropriate end-to-end reference model and architecture should be intensively investigated for such complex systems.

6.6.6 Coordinated APIs

In 5G mobile networks, it may be useful to define API so that applications and services can program network functions directly bypassing control and management to optimize the performance, e.g., to achieve ultra-low latency applications.

Discussions on the capabilities of the programmable interface should be objective-based, for example, accommodating a variety of application services easily, enabling the higher velocity of service deployment and operation, and the efficient physical resource utilization.

The users or developers who utilize the APIs could be categorized according to their roles. Application service providers enable value added services over the end-to-end virtual connectivity through the APIs. Advanced network service providers add some sophisticated functions to communications sessions, such as security and reliability, in order to facilitate faster application service deployment by the aforementioned application service providers. Network management operators also utilize the APIs for more efficient and agile resource handling.

Information modelling should be the most significant issues for the APIs development. It should include virtual resource characteristics, relationship between various resources, operational models, and so on. Levels of abstraction should be carefully investigated, so that the model and APIs should be human-readable and machine/system-implementable at higher performance simultaneously.

Since the considerations on software development methodologies would have the impact on the model development, the choice of the proper methodology for each capability will be important.

The system control and coordination architecture is another issue for the achievement of scalable and agile APIs. Not only the traditional provisioning/configuration or distributed control of networking systems, automatic and autonomic system control should be the main target of these activities. The closed loop control architecture might be the most innovative enhancement from the traditional networking systems even for the APIs.

The robustness and fault tolerance are absolutely necessary for the open systems controlled through the APIs by various providers. Isolation over the virtual resources should be carefully structured with APIs' functionalities and constraints.

In summary, discussions on the programmable interface capabilities should embrace;

- Level of abstraction sufficient both for system operations and for customization of the capability provided by the interfaces
- Modelling for the virtual/abstracted resource in a multiple-technology environment
- Ease of programming for service and operation velocity
- Technologies for automatic and/or autonomic operations
- Provisioning of classified functional elements suitable for a range of system developers such as supplication service providers, network service providers, and network management operator

Gap analysis

In 5G mobile networks, it may be useful to define API so that applications and services can program network functions directly bypassing control and management to optimize the performance, e.g., to achieve ultra-low latency applications. Information modelling should be the most significant issues for the APIs development. It should include virtual resource characteristics, relationship between various resources, operational models, and so on.

Discussions on the programmable interface capabilities should include:

- Level of abstraction sufficient both for system operations and for customization of the capability provided by the interfaces
- Modelling for the virtual/abstracted resource in a multiple-technology environment
- Ease of programming for service and operation velocity
- Technologies for automatic and/or autonomic operations
- Provisioning of classified functional elements suitable for a range of system developers such as supplication service providers, network service providers, and network management operator

These issues should be considered as a gap to be discussed for possible standardization items.

There is also a gap for requiring a completely new API to deal with new usage of softwarized network infrastructure such as capability exposure (refer to Clause 15).

6.7 Energy management aspects of network softwarization

Energy saving, optimization and management in the 5G networking and servicing ecosystem is an important issue in the design of the 5G infrastructure. As performance of 5G network equipment improves due to denser implementation a higher energy consumption would need to be avoided. 5G Network softwarization offers the means of flexibly and efficiently changing the configuration of the software components and network slices in a 5G systems for managing and optimizing the overall energy consumption in a multi-domain operation. Such energy management capabilities would help also in the use scenarios in other sectors which are using 5G systems. Network softwarization must

become a useful tool for reducing the environmental impact of other sectors and the means of controlling and managing energy consumption in the 5G infrastructure.

Network softwarization would enable monitor, and measure energy consumption and enable seamless, autonomic composition / decomposition of network slices, dynamic placements of network functions and migration of groups VMs [18]between servers of the same domain or across a group of energy-conscious domains aiming to i) optimize the overall energy consumption by dynamically changing the percentage of active versus stand-by servers/network functions and the load per active server in a domain, and ii) stabilize the 5G system energy distribution, under peak load and increased demand, by dynamically changing the energy consumption/production requirements of the local domains. Autonomic composition / decomposition of network slices, dynamic placements of network functions and moving VMs between servers in geographically distributed group domains is not trivial, as very strict Service Level Agreements (SLAs) should be guaranteed. Moreover, changing software components and/or containers for functions requires a significant telecommunication and energy cost, which should be precisely calculated.

Gap analysis

Energy-conscious 5G domain: optimizing the energy consumption within the limits of a single domain, based on system virtualization and the optimal distribution of VMs as well as M2M scenario. This will be coupled with the dynamic adaptation of active and stand-by servers/network functions and the load optimization per active server. A new monitoring framework to measure the energy consumption per server module/networking component and activate low-power states on devices would be needed.

Group of energy-conscious 5G domains: optimizing the cumulative energy consumption in a group of domains, based on optimal distribution of VMs across all of the servers that belong in the group of domains using policy-based methods. Measuring the energy consumption on the domain level and deploy policies and solutions that will achieve decreased cumulative power consumption across the whole group of domains would be needed.

6.8 Economic incentives aspects of network softwarization

The architecture design, resulting implementation and operation of 5G network softwarization are recommended to provide a sustainable competition environment considering social and economic issues. This includes drastic reduction of the components and systems lifecycle and operational costs for efficient and sustainable deployment, enabling appropriate return for all actors involved in the networking and servicing ecosystem, a reduction of barriers to entry in the 5G networking and servicing ecosystem and solving tussles among the range of participants in the ecosystem—such as users, various providers, governments, and IPR holders — by providing proper economic motivation.

A number of technologies have failed to flourish, or be justifiable due to inadequate or unsuitable architecture, concerning essential economic and social aspects (i.e. including contention among participants and/or lack of competing technologies, and/or lack of mechanisms to stimulate fair competition)

Gap analysis

Sufficient attention needs to be paid to economic and social aspects such as economic incentives in designing and implementing the 5G Network softwarization and its architecture in order to provide a sustainable competition environment to the various participants.

---

[18] Note that there exist a variety of technologies not only expressed as virtual machines (VMs) but various types of virtualization techniques such as resource containers, host-based virtualization, and hyper-visor based virtualization, etc. to serve for the same purpose.

Drastic reduction of the operational and lifecycle costs for all components and systems involved in the 5G network softwarization would be recommended for efficient and sustainable deployment enabling appropriate return for all involved actors.

Ways of resolving economic conflicts including tussles in 5G networking and servicing ecosystem that include economic reward for each participant's contribution are becoming essential. Different participants may pursue conflicting interests, which could lead to conflict over the overall multi-domain operation of Network softwarization and controversy in international/domestic regulation issues.

7 Integrated network management and orchestration

There are two aspects to consider for the integrated network management and orchestration for the IMT-2020 network softwarization. First aspect is how to manage and orchestrate the softwarized network components in a uniform and e2e way. The second one is how to softwarize network management and orchestration functionality.

The former should address traditional management functionality such as fault, configuration, performance, accounting, and security and orchestration functionality for a multiple domains, layers, and technologies within a single slice and for a multiple slices. Especially, orchestration among multiple slices if those slices belong to the same administration responsibility is a very challenging problem which hasn't been addressed up to now. ETSI NFV MANO architecture deals with orchestration of a single slice.

The latter should address optimal methodology of softwarizing and deploying management and orchestration functionality. The management and orchestration functionality consist of various sub-components such manager, agent, collector, virtual analytics instrumentation, autonomic management policy decision element, management information repository, etc. These components can be deployed as a standalone device or virtual function embedded in a hosting server. Optimal placement of these components is a new area to tackle with. In particular, dynamic runtime placement of such components is possible in the softwarized network environment to achieve optimal resource usage.

Another important issue is a multi-domain/layer/technology/slice orchestration. Within a single slice, there can be multiple domains of VNFs ranging from mobile access, core, and SDDC. Each domain can have one or more domain specific orchestrators. And some domain such as mobile core can have multi-layer (e.g., packet, PTN, OTN layers) networks and multi-layer orchestrator is required. For end-to-end orchestration, a multi-technology orchestrator that coordinates among domain or layer specific orchestrators is required. Also multi-slice orchestration is required for multiple slicesthat belong to the same network provider.

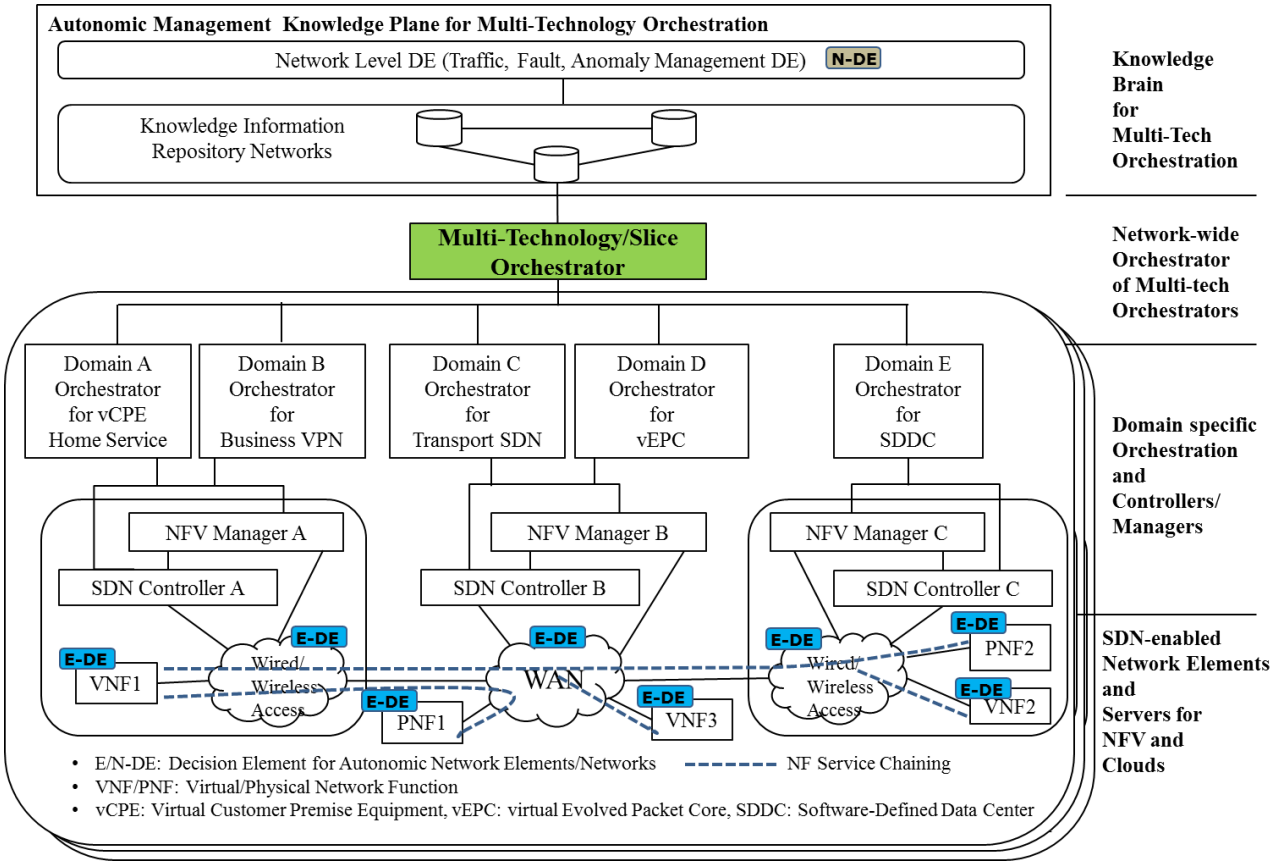


Figure 4 Multi-domain/layer/technology/slice Orchestration High-level Architecture

Figure 4 illustrates details of the multi-domain/layer/technology/slice orchestration. Within a single slice, there can be multiple domains of VNFs ranging from mobile access, core, and SDDC. Each domain can have one or more domain specific orchestrators. And some domain such as mobile core can have multi-layer (e.g., packet, PTN, OTN layers) networks and multi-layer orchestrator is required. For end-to-end orchestration, a multi-technology orchestrator that coordinates among domain or layer specific orchestrators is required. Also multi-slice orchestration is required for multiple slices which belong to the same network provider. Fig.x also shows that autonomic knowledge plane addresses optimal orchestration policy decisions.

Gap analysis:

There are two aspects to consider for the network management and orchestration for the network softwarization. First aspect is how to manage and orchestrate the softwarized network components. The second one is how to softwarize network management and orchestration functionality. The current technology gaps to be filled in are provided.

Some identified gaps which need to be filled in 5G network management are:

- Multi-technology orchestration across heterogeneous 5G multiple domains ranging from radio access, core, to SDDC.
- Multi-slice orchestration in case that multiple slices are allocated into a single tenant and service.
- Autonomic management capability for the efficient multi-technology/domain orchestration policy decision making.

- Optimal placement capability of virtualized management functional components in the 5G networks.
- Lightweight management and orchestration architecture to cope with performance problem and high OPEX
- In-system management capabilities empowering the network to control complexity using decentralization, self-organization, autonomy, and autonomicity as its basic enabling concepts. As such the management tasks are embedded in the network and the network as a managed system now executes management functions on its own.
- None-slio management and orchestration architecture to reduce both OPEX and CAPEX.

8 Mobile edge computing

8.1 General description

As we are approaching year 2020, new network service applications are emerging endlessly. While they may bring to the end user amazing experiences, they also require a more efficient, personalized, intelligent, reliable and flexible network.

Many OTT application providers have identified the demand of managing data at mobile edge which has significant advantages. OTT application providers will be able to access to the real time network context information so that it can timely adjust its traffic transmission. It will also benefit some OTT applications running in the cloud with locally processing of huge amount of data at mobile edge, the data of which is only valuable for just several seconds and which doesn't have to be sent to the cloud. Mobile users will be able to enjoy the personalized service with ultra low latency and higher bandwidth.

Nowadays, operator's key role is to maintain an efficient bearing network, which includes core network, radio network, radio fronthaul/backhaul network and backbone network, and the investment and maintenance of them, especially radio nodes (e.g. base stations and eNBs) and radio backhaul, are quite costly. Handling data traffic at mobile edge with providing network context to OTT applications will not only help operator explore new business opportunities but also can reduce radio and mobile backhaul resource consumption.

With the demands of all stakeholders, the concept of Mobile Edge Computing is raised in the industry. Mobile Edge Computing is an open IT service environment at a location considered to be the most lucrative point in the mobile network, the Radio Access Network (RAN) edge, characterized by proximity, ultra-low latency and high bandwidth. This environment will offer cloud computing capabilities as well as exposure to real-time radio network and context information. Users of interactive and delay-sensitive applications, which is located in proximity of the user, will benefit from the increased responsiveness of the edge as well as from maximized speed and interactivity.

IT economies of scale can be leveraged in a way that will allow proximity, context, agility and speed to be used for wider innovation that can be translated into unique value and revenue generation. All players in the new value-chain will benefit from closer cooperation, while assuming complementary and profitable roles within their respective business models.

8.2 Use cases and scenarios

Mobile Edge Computing technology enables a lot of new features in the mobile network.

- **Consumer-oriented services**: these are innovative services that generally benefit directly the end-user, i.e. the user using the UE, which includes gaming, remote desktop applications, augmented and assisted reality, cognitive assistance, etc. See 8.2.1 an example of consumer-oriented service.

- Operator and third party services: these are innovative services that take advantage of computing and storage facilities close to the edge of the operator's network. They are usually not directly benefiting the end-user, but can be operated in conjunction with third-party service companies, for example: active device location tracking, big data, security, safety, enterprise services, and etc. See 8.2.2 an example of operator and third party services.

- Network performance and QoE improvements: these services are generally aimed at improving performance of the network, either via application-specific or generic improvements. The user experience is generally improved, but these are not new services provided to the end-user. These include content/DNS caching, performance optimisation, video optimisation, etc. See 8.2.3 an example of network performance and QoE improvements use case.

8.2.1 Augmented reality

Augmented reality allows users to have additional information from their environment by performing an analysis of their surroundings, deriving the semantics of the scene, augment it with additional knowledge provided by databases, and feed it back to the user within a very short time. Therefore, it requires low latency and computing/storage either at the mobile edge or on the device.

In augmented reality services, UE can choose to offload part of the device computational load to a mobile edge application running on a mobile edge platform. UE needs to be connected to an instance of a specific application running on the mobile edge computing platform which can fulfil latency requirements of the application, and the interaction between the user and the application needs to be personalized, and continuity of the service needs to be maintained as the user moves around.

8.2.2 Data analytics

Some data analytic services need gathering of huge amounts of data (e.g. video, sensor information, etc.) from devices analyzed through a certain amount of processing to extract meaningful information before being sent towards central servers.

In order to support the constraints of the operator or the third party requesting the service, the applications might have to be run on all requested locations, such as mobile edge servers which is very close to the radio nodes. The application running on mobile edge server processes the information and extracts the valuable metadata, which it sends to a central server. A subset of the data might be stored locally for a certain period for later cross-check verification.

8.2.3 Mobile video delivery optimization using throughput guidance for TCP

Media delivery is nowadays usually done via HTTP streaming which in turn is based on the Transmission Control Protocol (TCP). The behaviour of TCP, which assumes that network congestion is the primary cause for packet loss and high delay, can lead to the inefficient use of a cellular network's resources and degrade application performance and user experience. The root cause for this inefficiency lies in the fact that TCP has difficulty adapting to rapidly varying network conditions. In cellular networks, the bandwidth available for a TCP flow can vary by an order of magnitude within a few seconds due to changes in the underlying radio channel conditions, caused by the movement of devices, as well as changes in system load when other devices enter and leave the network.

In this use case, a radio analytics Mobile edge application, which uses services of Mobile Edge Computing, provides a suitably equipped backend video server with a near real-time indication on the throughput estimated to be available at the radio downlink interface in the next time instant. The video server can use this information to assist TCP congestion control decisions. With this additional information, TCP does not need to overload the network when probing for available resources, nor does it need to rely on heuristics to reduce its sending rate after a congestion episode.

8.3 Key challenges

Mobile Edge Computing uses a virtualisation platform for applications running at the mobile network edge. The Mobile edge platform provides a framework for providing services to applications it hosts, with a basic set of middleware services already defined, allowing these applications to have a rich interaction with the underlying network environment, especially to be aware of the radio network status so that appropriate handlings will be made to adapt to the underlying network environment. In addition, radio analytic is exposed to applications through standardized API. See Figure 5, below for an overview of MEC framework.

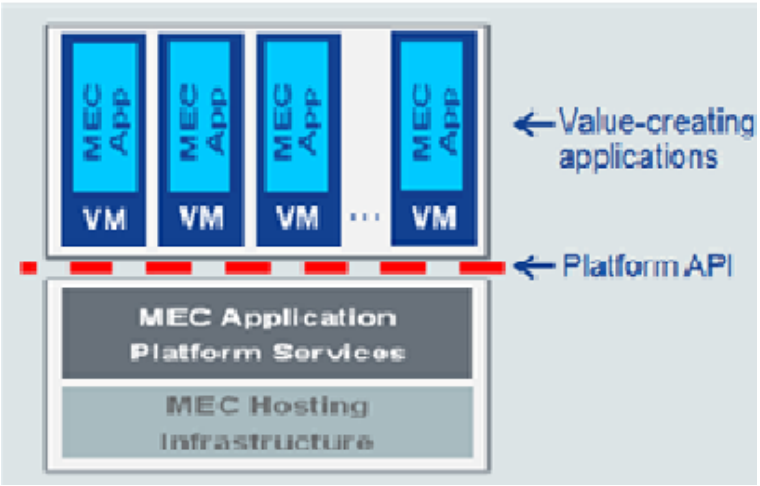


Figure 5 Overview of MEC framework

To achieve that, there are some key challenges to be considered:

8.3.1 Virtualization

Mobile Edge Computing uses a virtualisation platform for applications running at the mobile network edge. Network Functions Virtualisation (NFV) provides a virtualisation platform to network functions. The infrastructure that hosts their respective applications or network functions is quite similar.

In order to allow operators to benefit to as much as possible from their investment, it would be beneficial to reuse the infrastructure and infrastructure management of NFV to the largest extent possible, by hosting both VNFs (Virtual Network Functions) and Mobile edge applications on the same or similar infrastructure.

8.3.2 Mobility

Mobility is an essential component of mobile networks. Most devices connected to a mobile network are moving around within the mobile network, especially when located at cell edge, but also when changing RATs, etc, or during exceptional events.

Some mobile edge applications, notably in the category "consumer-oriented services", are specifically related to the user activity. It needs to maintain some application-specific user-related information that needs to be provided to the instance of that application running on another Mobile edge server. Therefore service continuity should be maintained while the user is moving to an area served by another mobile edge platform which hosts the application.

8.3.3 Simple and controllable APIs

In order to enable the development of a strong ecosystem for Mobile Edge Computing, it is very important to develop APIs that are as simple as possible and are directly answering the needs of

applications. To the extent this is possible, Mobile Edge Computing specifications need to reuse existing APIs that fulfil the requirements.

8.3.4 Application lifecycle management

The Mobile edge platform shall be available for the hosting of Mobile edge applications. The MEC management functionality shall support the instantiation and termination of an application on a Mobile edge server within the Mobile edge system when required by the operator or in response to a request by an authorized third-party.

8.3.5 Platform service management

The Mobile edge platform provides services that can be consumed by authorized applications. Applications should be authenticated and authorized to access the services. The services announce their availability when they are ready to use, and mobile edge applications can discover the available services.

8.3.6 Traffic routing

The mobile edge platform routes selected uplink and/or downlink user plane traffic between the network and authorized applications and between authorized applications. One or more applications might be selected for the user plane traffic to route through with a predefined order. The selection and routing during traffic redirection are based on re-direction rules defined by the operator per application flow. The selected authorized applications can modify and shape user plane traffic.

8.3.7 Data forwarding to edge or conventional computing server

User data needs to be categorized in two types depending on the service nature. One type is the data which are processed in application server of data center (DC) or the cloud. Another one is the service data which should be processed near the edge. For example, delay critical application data or localized proximity service data should be processed in the edge network, while some other application data are addressed to the conventional servers in DC or cloud. In order to conduct that way systematically, an identifier presenting data types and the control entity will be required in order to address the application data to edge network or to the conventional network.

8.3.8 Control signal transfer management

Because some type of user application data should be processed in the edge network, the service specific control signals may need to be combined with the edge local operation, or need to be transferred to the edge network for processing the control signals efficiently. Hence, a management capability will be required so that the control signals are combined or transferred to the local edge control entity for processing MEC application data.

8.3.9 Inter-edge mobility

Mobile edge service areas may consist of contiguous spots or isolated spots. A question arises how those proximity services can be seamlessly provided even when the devices move around the local areas across multiple edge networks. As an expected solution, such a capability will be required, that transfers the cached service data from a source edge to a destination edge server, and the Device Positioning Information will be useful to be shared among neighbor edge sites for tracking the mobile device, especially in the case that pin-point serving spots are distributed. That capability may be realized by means of some positioning systems or any spot marking assistance technologies.

Gap analysis

Support enhanced MEC management of virtualization

Mobile Edge Computing uses a virtualisation platform for applications running at the mobile network edge. Although Mobile edge server lifecycle management supported by existing NFV-MANO, while MEC management should support some enhancements in following aspects:

3) Mobile edge application lifecycle management: The MEC management functionality should support the instantiation and termination of an application on a Mobile edge server within the Mobile edge system when required by the operator or in response to a request by an authorised third-party.

Mobile edge application service management: The Mobile edge platform provides services that can be consumed by authorised applications. Applications should be authenticated and authorised to access the services. The services announce their availability when they are ready to use, and mobile edge applications can discover the available services.

Support inter-edge mobility

Mobility is an essential component of mobile networks. Considering some mobile edge applications are specifically related to the user activity, it needs to maintain some application-specific user-related information that needs to be provided to the instance of that application running on another Mobile edge server. Therefore service continuity should be maintained while the user is moving to an area served by another mobile edge platform which hosts the application. So MEC system should to support inter-edge mobility mechanism for service continuity

Support more simple and controllable APIs

In order to enable the development of a strong ecosystem for Mobile Edge Computing, it is important to develop APIs that are as simple as possible and are directly meeting the needs of applications. In addition, Radio Analytics/Radio Network Information is provided through standardized API and if there are enhancements required. MEC system should optimized existing APIs to make it more simple and controllable.

Support traffic routing among multiple applications

The mobile edge platform routes selected uplink and/or downlink user plane traffic between the network and authorized applications and between authorized applications. Morethan one applications might be selected for the user plane traffic to route through properly (e.g. video optimization, Augumented reality). The MEC system should support traffic routing mechanism among multiple applications: selection and routing during traffic redirection based on re-direction rules which is defined by the operator per application flow, and selected authorized applications can modify and shape user plane traffic.

9 Distributed cloud for service providers

9.1 Introduction

Cloud computing services are typically delivered via datacenter-like infrastructure. There are numerous types of datacenters with their respective advantages and disadvantages.

The Figure 6 depicts different types of datacenter:

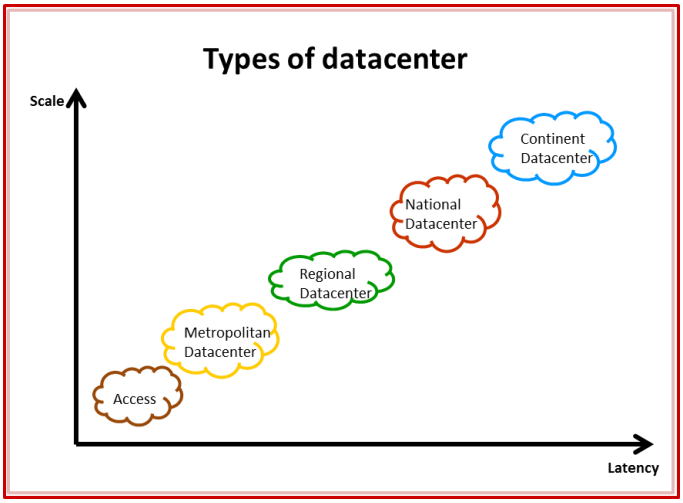


Figure 6 Type of datacenter

Various 5G applications have different requirements in terms of latency, throughput, and cost. Latency sensitive applications such as gaming, voice, video, and telepresence can benefit from the characteristics offered by regional, metropolitan, or access clouds. In a similar fashion, best-effort none-latency sensitive applications can profit from the cost advantage associated with national or continent-wide datacenter solution.

We can expect that in 5G some operators will leverage the distributed nature of their communication network by highlighting the benefits provided by regional or metropolitan clouds.[19]

A distributed cloud infrastructure supports the deployment and migration of applications between any of these types of cloud infrastructure. The capability to deploy the same application in either a national datacenter or near the network edge can bring unique advantages to operators. As example, a stock trading application would certainly gain from much shorter latency provided by a metropolitan-area rather than a centralized national cloud infrastructure.

A telecom cloud system management in 5G logically sits above the various types of datacenters. Such cloud manager decides in which particular site and processor pool an application would be initiated to fulfill the characteristics and cost defined in the Service Level Agreement associated with it.

A distributed cloud concept which encompasses processing and network resources is fundamental to the success of IMT-2020 network.

9.2 Key characteristics

In order to facilitate management of the system, both regarding traditional, manual O&M, and dynamic, automatic cloud orchestration as we see it in modern cloud based systems, the world of a VM is expected to be flat. I.e., we want VMs to be seen as running on top of a vast, uniform infrastructure where they can be moved around freely. This means that the physical infrastructure must be abstracted so that it can be presented to the VMs this way, as exemplified in the following Figure 7.

[19] P.Suthar, M.Stolic, "Building Carrier Grade TelcoCloud", IEEE, APWiMob,Bandung, August 2015

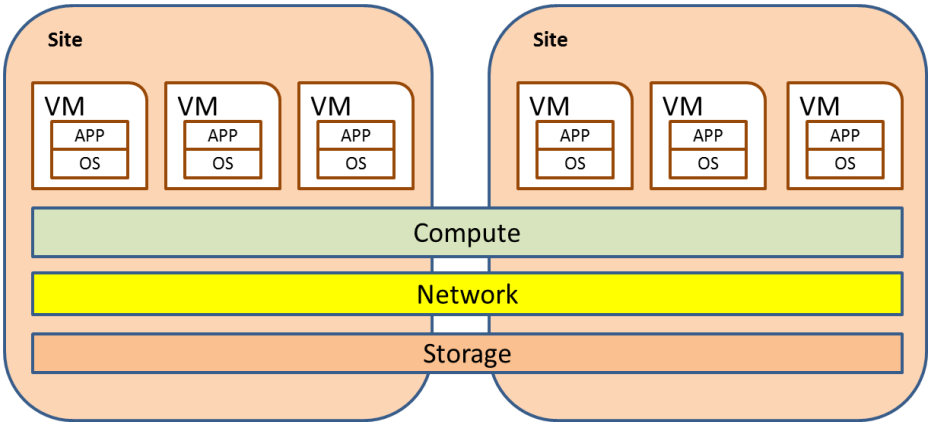


Figure 7 distributed cloud

The key characteristics of distributed cloud are:

- On-demand self-service

A consumer can unilaterally provision capabilities, such as server time, network storage and network bandwidth, as needed without requiring human interaction with each service's provider.

- Ubiquitous network access

Capabilities are available over the network and accessed through standard mechanisms that promote use by heterogeneous thin or thick client platforms (e.g., mobile phones, laptops, and PDAs).

- Location independent resource pooling

The provider's computing resources are pooled to serve all consumers using a multi-tenant model, with different physical and virtual resources dynamically assigned and reassigned according to consumer demand. The customer generally has no control or knowledge over the exact location of the provided resources. Examples of resources include storage, processing, memory, network bandwidth, and virtual machines.

- Rapid elasticity

Capabilities can be rapidly and elastically provisioned to quickly scale up and rapidly released to quickly scale down. To the consumer, the capabilities available for rent often appear to be infinite and can be purchased in any quantity at any time.

- Measured Service

Resource usage can be monitored, controlled, and reported providing transparency for both the provider and user.

In addition to these characteristics, the distributed cloud provides higher resilience and availability, greater security and the support of differentiating service classes, which will drive the deployment in order to guarantee the described Operational Level.

Gap analysis

The existing IMT network has its limitation and lacks of flexibility and agility for deploying the network functions and applications at any location where the performance and user experience would be optimized. In order to meet the extremely various demands of the services in 5G, for example, from ultra-low latency to high-latency tolerable service, distributed cloud technology provides a viable solution. To realize its benefits, the following gaps would need to be filled: (1) Distributed Storage Services that provide uniform, system-wide, distributed block storage for the applications (e.g., OpenStack's Swift and Cinder subsystems) (2) Networking Services that SDN enables such as

cloud-wide, virtualized connectivity, both at L2 and L3 levels, such as OpenStack's Quantum, (3) Distributed Compute Services that manage VM's, doings tasks such as start, stop, migration and supervisions of VMs is to be performed by a cloud computing service (e.g., CloudStack and OpenStack's Nova) and (4) Cloud management API for applications on top of the cloud infrastructure (e.g., OpenStack) for application deployment, migration and portability.

Although it is ideal to have all applications running inside VM's, reality, at least in the short term, dictates that some tasks must continue to execute on non-virtualized or specialized hardware. In order to limit the extra OPEX burden such system anomalies represent, it is still necessary to provide these environments with an API that makes it possible to manage them the same way (i.e. by abstracting and presenting a uniform interface to the applications) as is done with VM's, so that it is still possible to keep parts of the management APIs (loading, start, stop etc.) uniform and identical to the ones as in VMs.

10 In-network data processing

In-network data processing is a system that provides with network wide data processing and application services by network nodes. Increase of video traffic, the expansion of IoT, and shorter response time requirement need a basic structural change of current ICT system configuration where the data processing is done at the remote data center and the network just functions as a data pipe. In 5G network, it is required that the network node will provide with data processing and application services with the aim of reducing the network congestion and also shortening response time, when appropriate. ICN and the edge computing are typical examples. Edge computing works very well to shorten the response time and reduce the network congestion when the target data for computing is closed to an edge node area. In IoT, however, there are cases that the target data for processing span to many edge node areas, therefore the inner node of a network is more appropriate for processing. Another example can be the on-path data processing, which applies a series of data processing in tandem manner on the transmission path, and is frequently used in big data processing. There also exist the service provisioning cases that inner network node is better suited to perform, where the service user is sparse and distributed to several edge nodes.

Gap analysis

One use case scenario of in-network data processing is included in ITU-T SG13 that deals with requirements and architecture with in-network data processing. However, only a limited number of use case scenarios are described for In Network Data Processing. Further discussion for viable in-network data processing for 5G mobile network is necessary.

11 Resource usage optimization

A collaborative redundancy reduction methodology in an end-to-end path of softwarized network for resource usage optimization is a new problem to tackle with.

Taking advantage of SDN control, a collaborative redundancy reduction methodology is a new virtualized network functionality that dynamically offloads computational operations and memory management tasks of de-duplication to the group of the software designed network virtual functions. As this methodology efficiently chains storage de-duplication and network redundancy elimination functions and virtualizes de-duplication processes, it achieves effective performance without introducing high processing and memory overhead.

Gap analysis:

A large portion of digital data is transferred repeatedly across networks and duplicated in storage systems, which costs excessive bandwidth, storage, energy, and operations. Thus, great effort has

been made in both areas of mobile and fixed networks and storage systems to mitigate the redundancies. However, due to the lack of the coordination capabilities, expensive procedures of C-H-I (Chunking, Hashing, and Indexing) are incurring recursively on the end-to-end path of data processing. Redundancy reduction methodology in an end-to-end path of softwarized network may be needed for resource usage optimization.

Some identified gaps in 5G network end-to-end path resource optimization are:

- The de-duplication ratio of the client-side data reduction technique can be inefficient due to the limited data set and the processing cost can be too high for a client with limited capacity. Server side data de-duplication approaches have been used in traditional storage systems, and they mainly differ in the granularity of units for de-duplication, such as data chunks, files, hybrid, and semantic granularity.
- Network domain data reduction techniques suffer from high processing time due to sliding fingerprinting at the routers and high memory overhead to save packets and indexes.
- The ICN/CCN aims to reduce latency by caching data packets toward receiving clients. In addition, ICN/CCN uses name based forwarding table that causes extra table lookup time and raises scalability issues.
- Content Delivery Networks (CDN) can also reduce redundant data traffic by preventing a long path to an origin server after locating files close to users.
- In summary, currently available data redundancy reduction processes are very expensive, mostly performed by using vendor specific special purpose middleboxes or by introducing disruptive functionality. Furthermore, the costly processes are designed and performed independently, i.e., redundantly.

12 Resource abstraction

As 5G services provided on a slice utilizes end-to-end resources to implement functional components, it is necessary to define unified abstraction of resource to adopt for the best practice and performance of network. The detailed information of the physical resource can be abstracted so that other systems, applications, services, or users can access the capabilities of the virtual resource by using abstracted interfaces. Therefore, it is necessary to define unified abstraction of resource to facilitate API, as followings:

- Technology-agnostic representation of underlying physical-layer resource: The Network Softwarization should support any existing and future technology to be compatible with each other and can get the optimum and fair access of the underlying physical layer resources. Hence, this will enable the network technologies-independence capability.
- End-to-end characteristics/requirements: In order to provide context-aware services with high quality of service (QoS) for end-to-end resources, end to end QoS requirements must be ensured for all network infrastructure i.e. WiFi, LTE, etc and wired network (including ICN) as well.
- Granularity of slices, e.g., application session/instance granularity: The granularity of services can be determined according to the application requirement.
- Characterization of slices: By checking available resources optimum resource allocation can be done by the Network Softwarization and the slice fusion is supposed to be supported as well. It is also necessary to keep end-to-end principle in each slice to balance performance and cost of the 5G system.

In short, resource abstraction purpose is to ensure transparent network technologies and architecture

for user. In other words, this 5G resource abstraction allows end-user to handle and use underlying resources in 5G network infrastructure easily (enable and maintain simplification for network users).

Gap analysis

This is no common model that can provide abstraction of various capabilities supported by physical resources that constitute end-to-end scope and are not covered by existing networks, including, physical radio interfaces, packet forwarding and routing in access networks. The granularity of current abstraction model may not be sufficient to support various approaches to satisfy end-to-end quality of application, while minimizing impact on utilization of networks.

13 Migration towards newly emerging networks

Implementation of network softwarization for 5G networks will co-exist with legacy network equipment and be compatible with the existing network technologies. In other words, it should work in a hybrid network composed of classical physical network appliances and network softwarization appliances during the deployment phase. Therefore, the migration from the starting network to target 5G network by Network Softwarization can gradually be done by using the hybrid network deployment, as following three-steps-migration path:

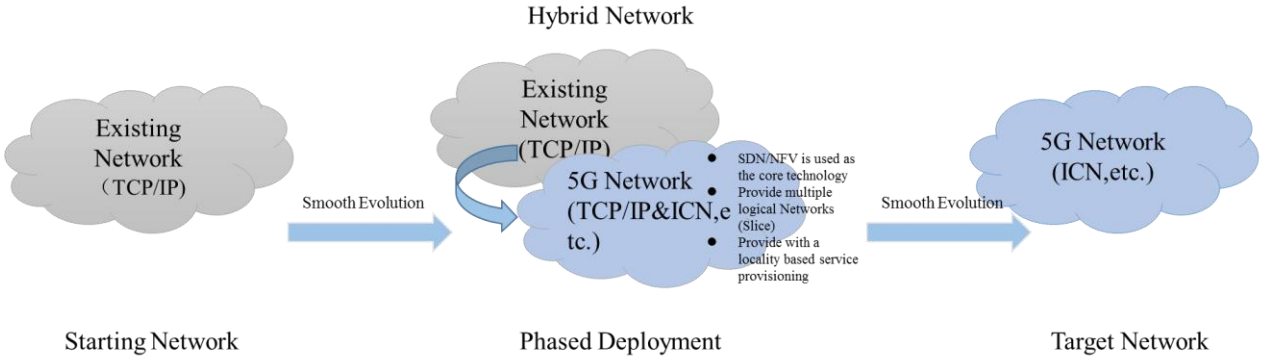


Figure 8 Draft Phased Migration

Starting network:

The starting network phase utilizes current and state-of-the-art network technologies (existing technologies): LTE, IP-based network, etc.

Phased deployment (intermediate phase):

The benefit of this deployment is during the migration intermediate phase, all end-to-end resources can still maintain the conventional communication means to communicate with each other. As a result, this mechanism enables migrated end-to-end resources are deployed in conjunction with existing devices. It enhances the migration process feasibility by enabling both the gradual 5G deployment and maintains current communication models simultaneously during intermediate period.

The requirements for this 5g network migration are stated as followings:

- The 5G network is a foundation of the future network and having a mechanism to smoothly evolve to the future network which is under discussion in ITU-T SG13/Q15:
- The migration scenario from the early stage of 5G network to the future network:
- A locality based service provisioning mechanism and architecture: Mobile Edge Computing, which is one of the hot topics of 5G mobile system discussion, and local area computing are

examples.

- Possibility of the in-network data processing/service provisioning capability, where each network node carries out some data processing and service provisioning: This feature is especially efficient to handle IoT and big data.
- Adoption of emerging network technology:

And the possible technological directions are also stated as followings.

- The application of network softwarization concept as a core technology to make 5G network: SDN and NFV are the examples.
- Adoption of multiple logical networks (Slice), each having different architecture that fits to the major services provided on the slice: IP slice, ICN slice, IoT slice, and low latency slice, etc. will be candidates.
- Having clear API for the development of a variety of applications and services developments and their provisioning:

Moreover, it is requested that the 5G network will provide with an in-network data processing capability, where each network node carries out some data processing and service provisioning. This feature is especially efficient to handle IoT and big data.

Target network:

In the final phase, the target network is formed and the 5G deployment is fully achieved (the migration process is finished), i.e. all the parts meet the requirements of network softwarization for IMT-2020 networks. Furthermore, in order to guarantee high QoS, low latency and high reliability of future network, the 5G deployment should be based on the state-of-the-art network framework, such as ICN which is referred as Data Aware Networking (DAN) in ITU documents (ITU-T Y.3033) and widely-recognized for its high performance and low latency.

Gap analysis

Network virtualization described in [ITU-T Y.3011] allows the network providers to integrate legacy support and keep backward compatibility by allocating the existing networks to LINPs (i.e., slices) for deploying new network technologies and services or migrating to new network architecture.)

It is expected that network softwarization, especially slices can provide migration paths to newly emerging network architecture since it may be possible to accommodate multiple network architectures in slices concurrently. However, there is yet no activity observed for discussing the detailed migration scenario.

14 RAN virtualization and slicing under software control

Mobile network virtualization can be also found in radio access network (RAN) as a fundamental network domain that can be realized by network softwarization as described in the previous clause. RAN virtualization may be found typically in a fabric of Cloud-RAN (C-RAN) structure as described below.

The 5G mobile network needs to support flexible network capabilities with software defined radio access technologies in terms of frequency band, transmission schemes, antenna configuration, multiplex access attributes, for example, in order to achieve network optimization for a variety of services in dynamic manner in various environments with a reasonable cost, In such circumstances, a flexible programming scheme with software controlled virtualization will be required for gaining benefits of network capabilities and performances for network operators, service providers and end users.

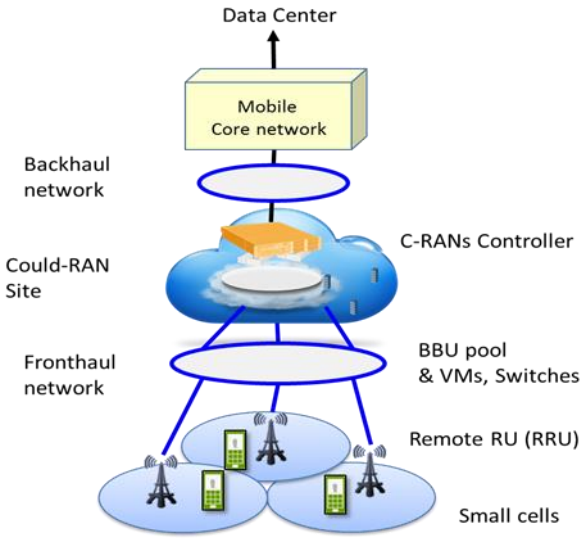


Figure 9. Virtual RAN physical model - example

A physical model of Cloud-RAN is generally illustrated in Figure 9 which consists of RAN control platform, BBU pools, Backhaul connection to Core network, Fronthaul connection to a number of small cell sites with remote radio units (RRU). In 5G novel network, the future RAN is expected to have an intelligent control over functions and transport networks in the cloud RAN, and some number of small cell sites with various radio network resources which can be controlled remotely from the central controller.

An overall picture around possible RAN virtualization is illustrated in Figure 10. It enables software based control of data transport from user device of access network, fronthaul, backhaul to core network.

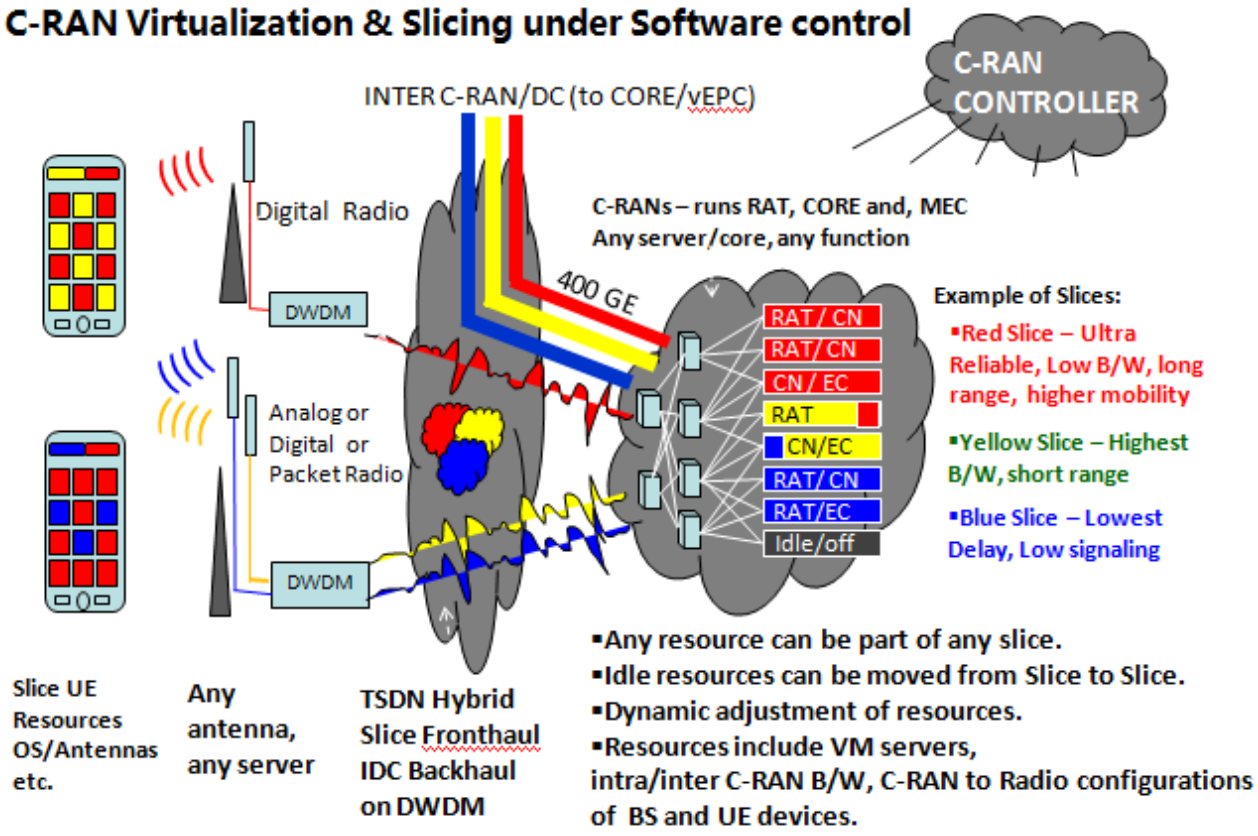


Figure 10. RAN Virtualization model - example

Concepts of the slice and RAN domain virtualization are introduced in this Figure 16. In this example, three slices are mapped in association with three types of service profile to achieve the required quality and reliability by means of network key specs such as bandwidth, propagation range, latency, mobility, UE configurations, and so on. Scope of the slicing can be extended to the network attributes of radio stations and user terminals and the data center (DC).

RAN virtualization may own essential capabilities as follows for increasing network gain:

- RAN controller integrates overall network control, scheduling, and data transport control throughout the user devices, remote radio unit (RRU), radio access schemes, fronthaul, backhaul, and radio resources such as transport bandwidth, RAT attributes, signaling on BBU,
- Depending on the requirements for service application, the network slices are flexibly arranged and scaled up or down in a configuration set with appropriate network resources, virtual network functionalities (VNFs) in the virtual network topology in dynamic way.
- The RAN has a capability of orchestrating the VNF chain by arranging and scheduling the virtual machines, storage memories, processing units, and so on. In consequence, all the data processing functions like the vEPC and the transport lines become programmable in software.
- Transport SDN is also a softwarized capability for the data transport control in the path of X-hauls through the appropriate virtual network functions on the slice per the necessary data services and performances. The fronthaul and backhaul network path, the bandwidth, QoS, wavelength, time-slots, etc. are controlled by the TSDN in the routing between the RRUs and the core network.

- On each slice, the network resources are flexibly allocated in a scalable manner under the RAN controller. Network resources are pooled, and idle resources are re-locatable among some network slices.

As a result of the expected capabilities as above, the network can provide comfortable quality of experience (QoE) for a variety of services in a reasonable cost (CAPEX and OPEX) with higher flexibility and agility. However, in order to realize it, some gaps need to be studied for overcome.

Gap Analysis:

Virtualization of RAN domain in conjunction with software control is expected as effective solution to provide appropriate QoE for diversified service requirements in dynamic way with a reasonable cost. RAN resources and the functionalities are mapped onto the network slices in association with the service profiles.

Following elements should be defined.

(1) Slice management and the arrangement of VNFs, virtual topology, software based transport control on the slice.

- Control of resources mapping to slices.
- Computation of resource to slice assignments, trade-offs.
- Decision of the timing when to turn on/off a slice.
- Control of slicing within the UE, OS etc.
- Management of the end to end resources on a given slice.
- Inter-slice management (moving resources between slices)

(2) Activation of slice attributes such as the application drive, resiliency, OAM, and security on each slice and inter-slice.

- Start/Stop/Management of applications residing within a slice.
- Resiliency of slice control
- OAM within a slice, between slices.
- Security within slice, between slices.

(3) Appropriate APIs in some network elements such as follows:

UE, Xhaul, TSDN, NVFs, Hypervisors, Switch/Routers, OTN/DWDM devices, clock synchronization, etc.

Reference documents on C-RAN

[ITU-R REP M.2320] Report ITU-R M.2320-0 (11/2014), “Future technology trends of terrestrial IMT systems”, 5.6.4 Cloud-RAN

[NGMN] “SUGGESTIONS ON POTENTIAL SOLUTIONS TO C-RAN BY NGMN ALLIANCE”, DATE: 03-JANUARY-2013, VERSION 4.0

[CMRI] China Mobile Research Institute, “C-RAN, The Road Towards Green RAN (Version 3.0)”, White Paper, Version 3.0 (Dec, 2013)

[ARIB] ARIB 2020 and Beyond Ad Hoc Group White Paper, “Mobile Communications Systems for 2020 and beyond”, Version 1.0.0, October 8, 2014, A.7.4 Cloud-RAN (C-RAN)

15 Capability exposure

Network capability exposure is for operators to provide the network capabilities needed by the 3rd party ISP/ICP. These service providers should be able to flexibly and efficiently use the capabilities via e.g. open API, while operators enhance the network function to provide the new capabilities. Network operators cooperate with the 3rd party to build an ecosystem to improve user's experience and benefit from some new business models.

15.1 Architecture

Through the three layer architecture of network capability exposure shown in Figure 11, the operator could offer the network capabilities to the third party users.

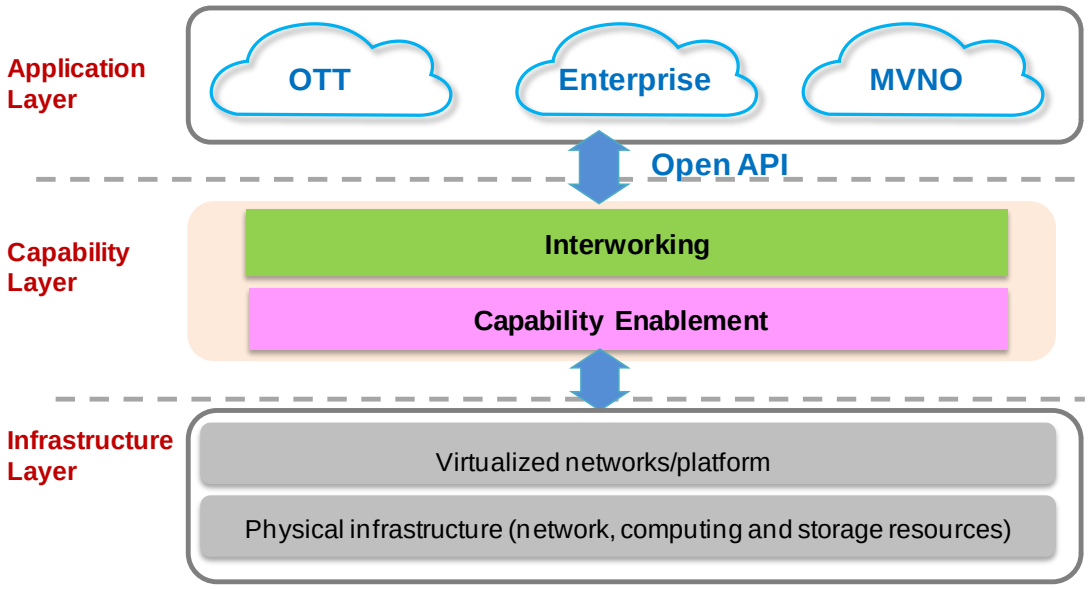


Figure 11. Architecture of Network Capability Exposure

Application layer: The third party platform or servers located at the highest layer are the consumers of the network capabilities which calls the APIs provided by capability layer. These capabilities may be to retrieve network state information, configure capability rules, apply for network control functions, build dedicated network slices, etc.

Capability layer: This layer is located between the application layer and the infrastructure resource layer. The primary function of this layer is to interwork with the 3rd party via open APIs to analyse the service requirements and to enable the infrastructure resource capability to meet the specific demand. For example, for the 3rd party with on-demand network purposes, the capability layer should orchestrates the network resource and builds dedicated network slices. Capability layer functions include:

- Interworking unit: The task of this unit is to receive the requirement and service information from the third party, then offer the required network capabilities to fulfil the interaction with the third party requirement.
- Capability enablement unit: this unit provides the map between the third party invocation and network capabilities. In addition, it exposes the information from the under layer network such as state data of control plane, user plane and service, value-added services, and infrastructure resources (computing, storage and connection).Moreover, it orchestrates the network resource according to the third party application requirement.

Infrastructure resource layer: It includes 5G network function comprising access nodes, network

functions and 5G devices(presented as (smart) phones, wearables, CPEs, machine-type modules, etc.), and the physical resources like cloud resources, network links, and etc.

Gap analysis

NGMN 5G White paper has proposed the requirements on network capability exposure to enable business agility, and envisioned 5G architecture that includes the network capability exposure as a key functionality.

4G network architecture enhancement for capability exposure is an ongoing study in 3GPP SA2. 5G capability exposure work is on the stage of service requirement research in 3GPP SA1. It mentions that the network slicing capability in 5G era could be exposed to the customer by providing it the specific network slice according to its demand. However, the detailed discussion on capability exposure is not yet done and current use cases are not comprehensive and systematic. The existing ITU-T specifications have covered some aspects in NGN context [ITU-T Y.2234 and Y.2240 from a capability perspective],

The following points should be studied:

- Scenarios and requirements of network capability exposure
- Architecture, mechanism and API of capability exposure
  - Overall architecture for the capability exposure
  - Potential solutions and E2E procedure to enable each capability to fulfill the specific service demand
  - Open APIs interworking with 3rd party based on the investigation on API work of this document.
  - Privacy and Security

16 Identification of gaps

Please refer to Clause 7.2 of the main body of the FG-IMT-2020 report. It contains the standardization gaps originally in this clause.

17 Supplementary material:

17.1 **Satellite Networks aspects of network softwarization**

Editor's note: this material was originally in an Annex of the document produce by the softwarization group within FG-IMT-2020.

Satellites have enabled since the beginning the extension of telecommunication services, exploiting their main attribute: ubiquity. Depending on the condition of static or mobile for the satellite terminals granting access to satellite telecommunication services, Fixed or Mobile Satellite Services are defined:

- Fixed-Satellite Services (defined primarily for GEO satellites and for broadcast, telephony and data communications) can cover large regions with limited resources and provide telecommunication services to millions of users simultaneously (i.e. broadcast of TV programs).

- Mobile Satellite Services (for GEO, MEO and LEO satellites) have been growing as the technology enabling their advances, and with the implications of addressing a market with strong dominance from terrestrial Mobile Network Operators (MNOs). It is in regions with limited or no telecommunications infrastructure (off-shore, rural environment or areas where the digital divide is most present) where MSS target their main penetration, as well as to provide with backup for terrestrial telecommunications networks in case of failure.

The introduction of IP into satellite networks dominates most of the functionalities and improvements being added, the objective behind having IP flowing equally through satellite networks as it does through terrestrial networks is clear: making of telecommunication satellites part of the existing telecommunications infrastructure despite satellite communications' particularities. Depending on the satellite orbit, the distance between transmit and receive stations and the satellites ranges from roughly 200 Km for Low Earth Orbit Satellites (LEOs), to slightly more than 35.000 Km for Geostationary Earth Orbit Satellites (GEO). Traditionally LEO and GEO satellites have been mostly used for telecommunication services, each with their own advantages when compared to the other:

For the provision of most services, satellite networks are always connected to an IP backbone. Even if not integrated in terms of offering full interfacing capabilities to procure NFV and SDN, we could say that in today's heterogeneous telecommunications infrastructure, satellite networks already play their role as part of the overall network infrastructure.

The main role for satellite communications can play is to extend **5G Cloud & networking s**ervices to remote, off-shore and rural locations, in cases where telecommunications' networks infrastructure is scarce. Additionally, satellite communications can provide an alternative pathway for congested terrestrial networks. With traffic set to increase monotonously throughout the foreseeable future, and the improvements in capacity procured by new GEO satellites as well as the reduction of overall latency brought by new satellite constellations orbiting closer to Cloud Service users **makes the case for the use of satellites as part of Could Networks infrastructure.**

In addition one directional broadcast satellites could provide a way of forwarding various forms of content to CDN/ICN nodes avoiding terrestrial networks and lessening congestion. The ability of High Throughput Satellites to release high capacity over differentiated areas of cover**age or spotbeams independently,** helps segregating the contents being delivered to diverse regions (CDN/ICN nodes).

<u>Gap analysis</u>

A key challenge is to develop a pragmatic and cost efficient oriented integration of satellite with terrestrial scenarios in the **5G networking ecosystems.**

An evaluation of the benefits of gracefully integrating satellite networks and cloud networking as an intrinsic part of the 5G ecosystem **would elicit the relevant network softwarization enablers for the use of satellites in 5G networks.** The **roadmap** and the methods for integrating satellite **in** terrestrial **5G** architecture **would represent an additional challenge.**

17.2 Legacy Data Reduction Techniques

Editor's note: this material was originally in an Appendix of the document produce by the softwarization group within FG-IMT-2020.

A number of data reduction techniques have been proposed for both networks and storage domains. Storage domain data reduction techniques are to save storage space, and run at a client's or at a server's side. Client side techniques can reduce the network bandwidth by eliminating redundancy before the data transfer.

Network domain data reduction techniques are to save bandwidth and reduce latency by reducing repeating transfers through network links. End-to-end Redundancy Elimination (EndRE) and WAN optimizers remove redundant network traffic at two end points (e.g., branch to head quarter). Network Redundancy Elimination (NRE) techniques eliminate repeating network traffic across network elements such as routers and switches. NRE computes indexes for the incoming packet payload and eliminates redundant packets by comparing indexes with the packets saved previously. The redundant payload is encoded by small sized shims and decoded before exiting networks.

The ICN/CCN aims to reduce latency by caching data packets toward receiving clients. In addition, ICN/CCN uses name based forwarding table that causes extra table lookup time and raises scalability issues. Meanwhile, Content Delivery Networks (CDN) can also reduce redundant data traffic by preventing a long path to an origin server after locating files close to users. Fig.x shows the various efforts of resource optimization by eliminating redundancies in specific ways.

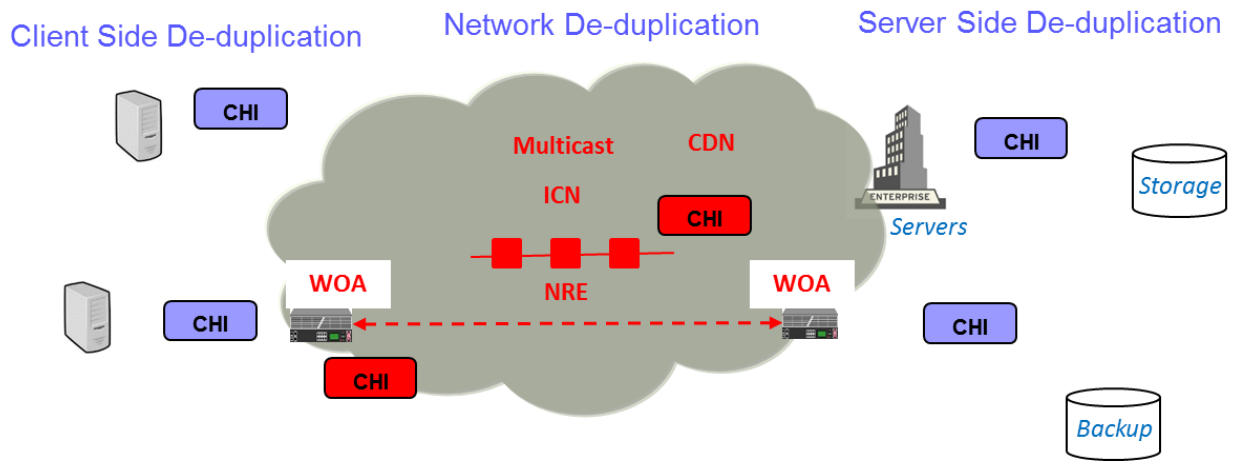


Fig.x. Legacy Redundancy Elimination Techniques

17.3 Standardization efforts in mobile edge computing

Editor's note: this material was originally in an Appendix of the document produce by the softwarization group within FG-IMT-2020.

17.3.1 Standardization efforts in the industry

The concept of Mobile Edge Computing has now been widely accepted in the industry and is implemented with many reference cases around the globe. Therefore, there's great need of standardization efforts to enable application portability and good performance in multi-vendor environment.

On September 24th 2014, a new multi-stakeholder industry initiative on Mobile-edge Computing (MEC) was formed under the auspices of an ETSI Industry Specification Group (ISG).

The purpose of the ISG MEC is to produce (Standards Track Deliverables) interoperable and deployable Group Specifications that will allow the hosting of third-party applications in a multi-vendor Mobile-edge Computing environment. The deliverables of ISG MEC include:

- an ontology containing the terminology that will be used consistently by the set of MEC specifications (informational)
- a gap analysis, identifying critical functional elements and techniques that need to be standardized to provide greater value (informational);
- Requirements (normative);
- Framework and reference architecture (normative);
- Specifications relating to the platform services and the APIs (normative).

In February 2015, 3GPP kicked off Feasibility Study on New Services and Markets Technology (SMARTER), the objective of which is to develop 5G technical requirements specification. Until SA#69 plenary meeting, there are 49 use cases defined, many of which are MEC relevant, such as Tactile Internet, Extreme real-time communications and the tactile internet, Improvement of network capabilities for vehicular case, Connected vehicles, Routing path optimization when server changes, etc.

In September 2015, 3GPP SA#69 has agreed that SA WG2, being responsible of developing 3GPP architecture standard, were asked to provide a proposed study item to TSG SA#70 (December 2015) on Next Generation Mobile Radio Technology. Requirements identified by SA1 SMARTER and other operator pain points will be addressed in the SA2 study item, and it can be speculated that MEC will be one important feature of it.

17.4 Supplemental figures on network softwarization

Editor's note: this material was originally in an Appendix of the document produce by the softwarization group within FG-IMT-2020.

17.4.1 IoT services

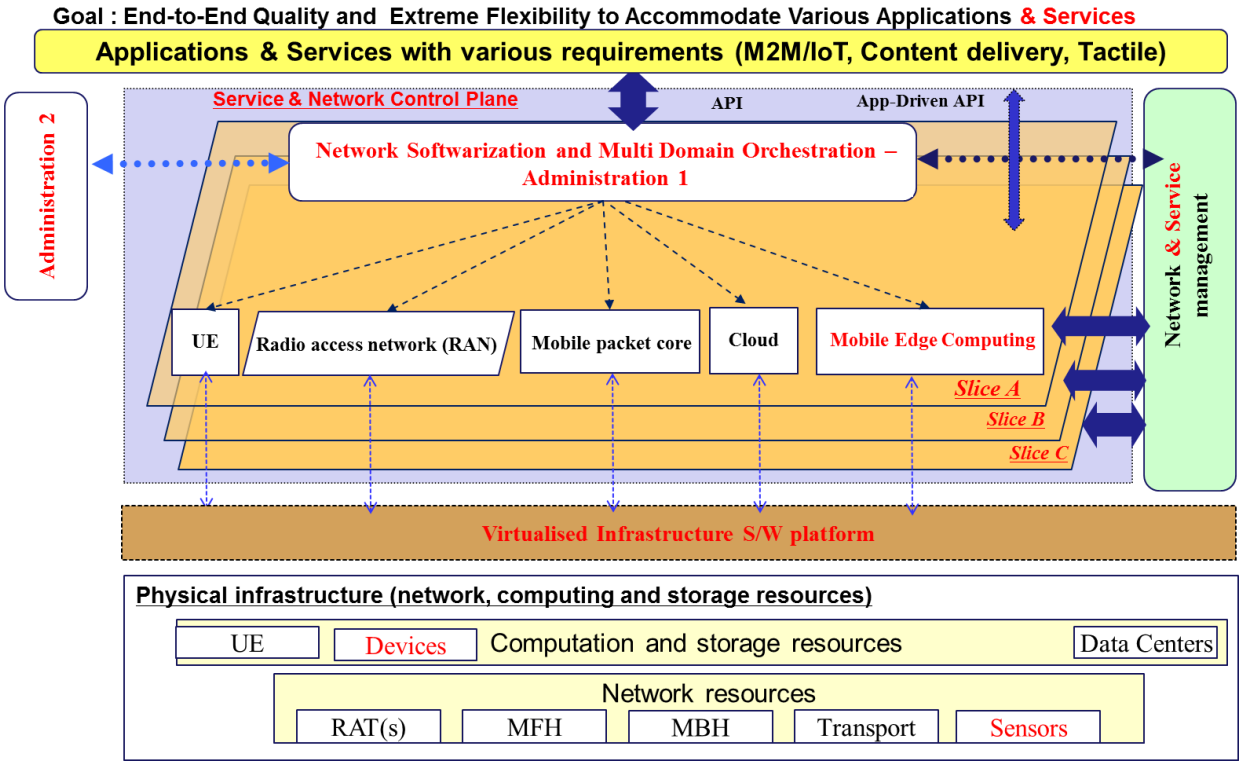


Figure III-1 A use case of using the architecture to implement IoT services

17.4.2 Transport service

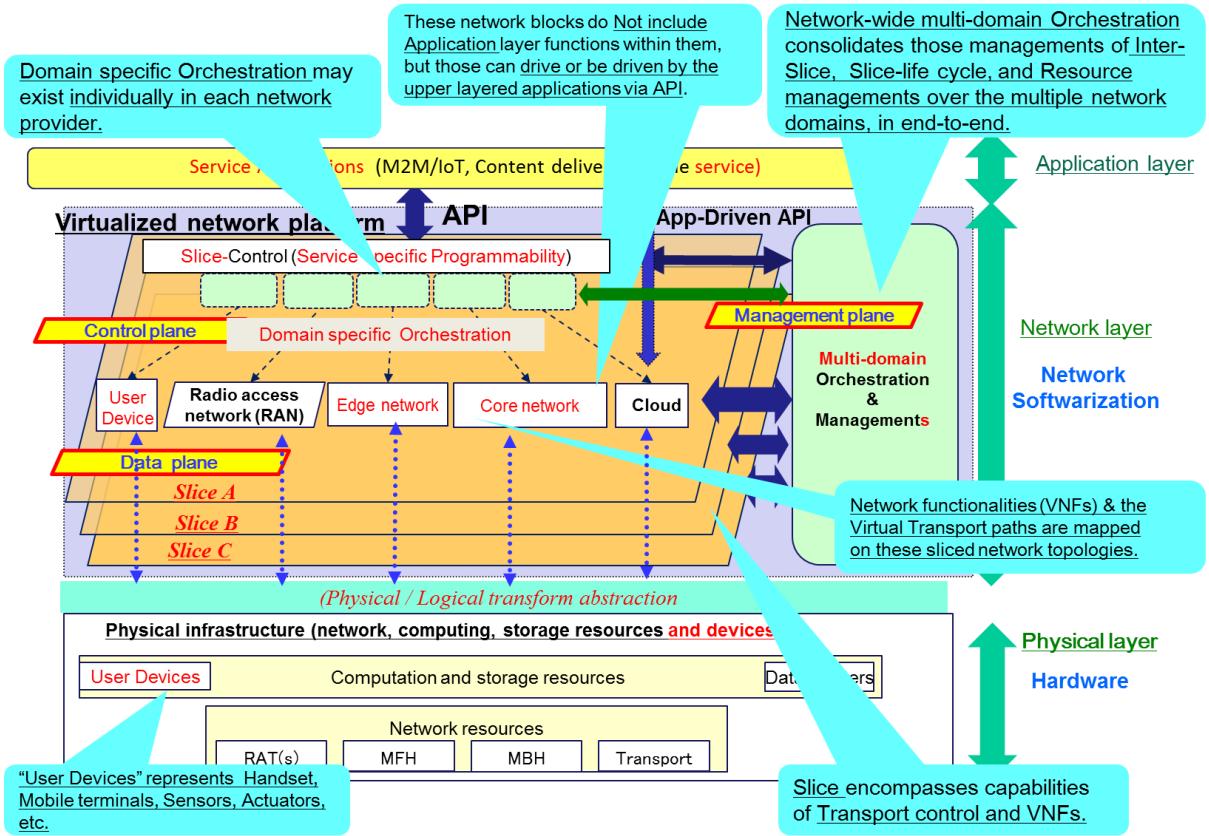


Figure III-2 A use case of using the architecture to implement transport services

Figure III-2 shows a use case of using the Figure 2 to implement transport services. The large part of the middle area represents the virtualised network platform that is the target of network softwarization in this use case.

The software platform consists of multiple slices on which a Slice-Control entity exists to conduct programming for the service specific control on each slice with the possible individual domain orchestrations. On each slice, virtualized network functions (VNFs) are placed appropriately in those virtual networks of RAN, Edge network and Core network throughout the end-to-end path to connect the user device to service cloud, while the data transport control is conducted.

These network blocks do not include application layer's function within them, but can drive or be driven by the upper layered application functions via API. And, the application data are carried through the logical transport path mapped logically on the slice network under the data forwarding control.

Those slice-controllers under individual domain are totally organized in end-to-end by the multi-domain orchestration which coordinates the inter-slice management, the life cycle management and the resource management.

17.4.3 Detailed functional architecture

Figure IV-3 describe a detailed functional architecture.

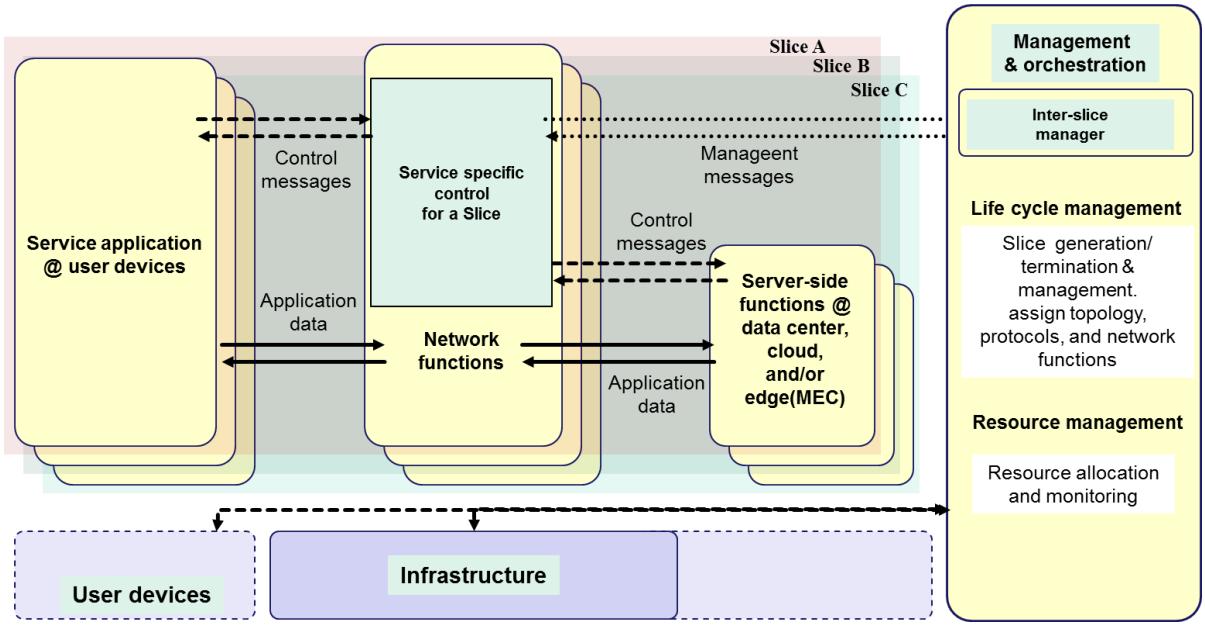


Figure III-3 Detailed functional architecture

Service-specific control for each application service are allocated on each slice. Different application services may have different control requirements which request different types of functions and

resources (physical and virtual) and topologies to be instantiated and different configurations to be maintained during their life time.

Inter-slice control coordinates service-specific controls for slices, and manages a common control functions in the Management and Orchestration block. It interfaces with service-specific controls to perform life cycle management and resource management of slices.

Note that while a service specific control may track authentication of its service application, there may be a case where a physical device be tracked by the management and orchestration block in some way. This is because a device may be connected to multiple slices simultaneously. Also note that server-side functions may be located on the infrastructure provided by other parties.

Management and Orchestration block is responsible for life cycle management of slices. It performs placement and instantiation of network functions. Furthermore, it performs association to the function on user devices and server-side functions.

At the time when a service-specific slice are to be created, requests may be generated by the service-specific control indicating what network functions are needed (e.g. any MTC service, CDN service, Public Safety) and what type of devices & applications (e.g. Video, Device data /Real-time or not) are used in their locations.

Management and Orchestration block is responsible for resource management of infrastructure. It manages the allocation of network functions and virtual networks which are used by slices. It examine the requests and determine the resources to be allocated, then it instantiates the network functions and virtual networks on the slice on associated physical infrastructure.

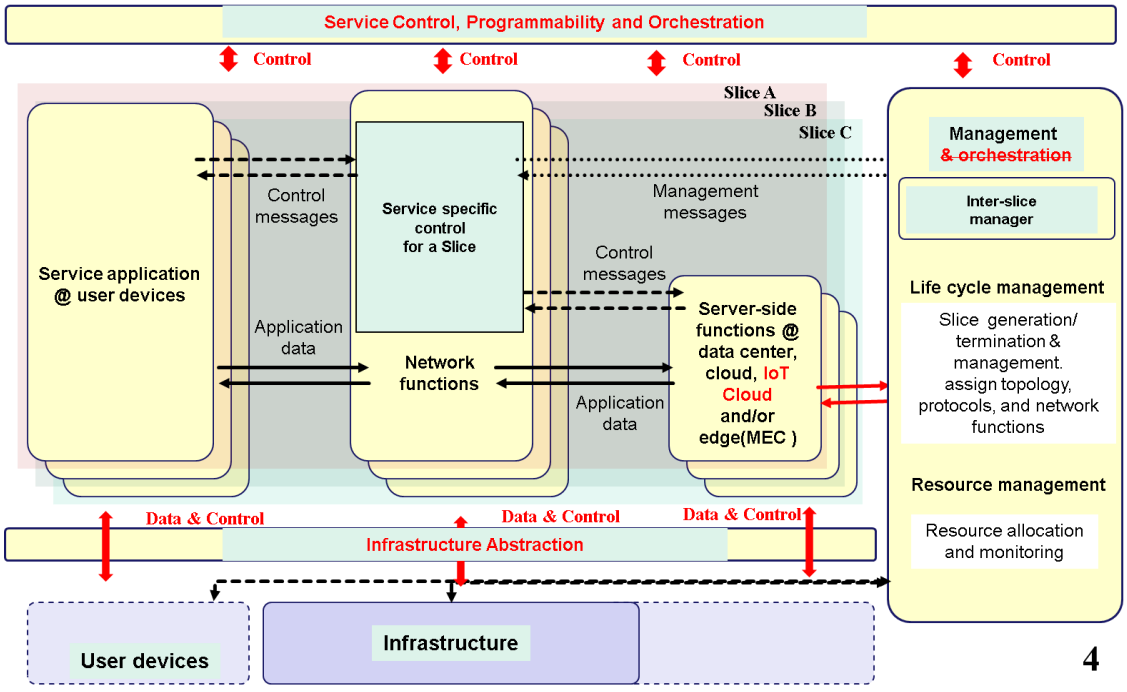


Figure III-4 A use case of the functional architecture to implement IoT services

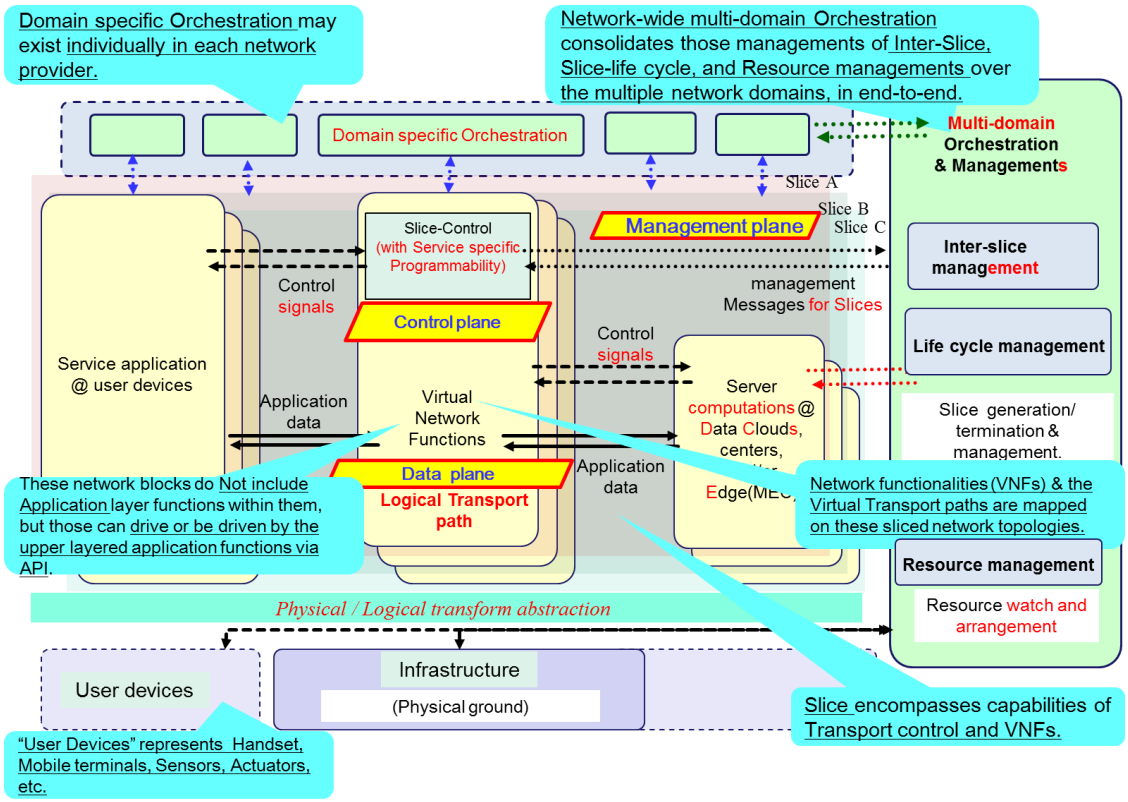


Figure III-5 A use case of the functional architecture to implement transport services

Figure IV-5 shows a functional model and the end-to-end signal/data flow of 5G mobile network. The multi-domain orchestration consolidates the domain specific orchestration operations in network-wide, and that conducts total managements of Inter-slice, Slice life cycle, and Resource management over the multi-network domains in end-to-end.

On the other hand, the domain specific orchestration may exist individually under each network provider and plays a role of virtual network organization on each network administrative domain. On each slice, the slice encompasses capabilities of logical transport network control and virtual network functionalities (VNFs), and the virtual transport paths are mapped on the sliced network topologies.

18 Acknowledgement

This work was partially supported by the European Union 7th Framework Program DOLFIN project ("Data centres optimization for energy-efficient and environmentally friendly internet"; http://www.dolfin-fp7.eu), the EU H2020 5G PPP projects: 5GEX ("5G Multi-Domain Exchange"; https://5g-ppp.eu/5GEx) and SONATA ("Service Programing and Orchestration for Virtualized Software Networks"; https://5g-ppp.eu/sonata/) and the European Space Agency (ESA) INSTINCT project ("Scenarios for integration of satellite components in future networks"; https://artes.esa.int/projects/instinct).

Contributors (in Alphabetical Order)

This is the list of all contributors who submitted any written form of comments or contributions.

- Hui CAI, China Mobile
- Wei CHEN, China Mobile
- Taesang CHOI, ETRI
- Ken FUJIMOTO, Huawei Technologies Japan
- Alex GALIS, University College London
- Sherry SHEN, Nokia Networks
- Marc MOSKO, PARC
- Akihiro NAKAO, University of Tokyo
- Takashi SHIMIZU, NTT
- Xiaowen SUN, China-Mobile
- Prakash SUTHAR, Cisco Systems
- Toshiaki SUZUKI, Hitachi
- Akihiko TAKASE, Hitachi
- Toshitaka TSUDA, Waseda University
- Jian WANG, Ericsson
- Weixin WANG, Nokia Networks

Appendix III

End-to-end QoS

Editor's Note: Appendix III was produced during the FG-IMT 2020 focus group in order to investigate gaps in standardization related to IMT-2020. While the request from SG-13 was to deliver a report outlining standardization gaps, the consensus of the focus group was that the working documents produced and used during the focus group work contained useful information for future work and should be captured. Note, however, the focus group concentrated on producing accurate descriptions of the standardization gaps in the main body of this document; some minor errors may exist in the appendices. They are, however, the output of the focus group but are provided for information only.

Editor's Note: This appendix uses clause references in a form usually associated for normative text. This is maintained for this report to align with references made in the main body of this report.

This baseline document for a QoS framework for IMT-2020 is revised on related input documents, as well as discussions, feedback and suggestions received at the break-out session on QoS on the last Torino meeting.

The revision is as follows:

- Addition of section 2, references
- Two editor's notes in relation to new parameters and measurement & monitoring
- Some minor editorial corrections

==========

QoS framework for IMT-2020

Summary

This deliverable:

- provides concepts and descriptions of Network Performance, Quality of Service and Quality of Experience
- illustrates how the concepts are applied in IMT-2020, including the relationship between concepts
- indicates and classifies concerns for which reference model and parameters can be needed
- identifies generic guidance of performance parameters, QoS classes and its allocation

Editor's Note: Depending on the developed contents of work items in the focus group (e.g., High-level Architecture, Network Softwarization and Use Case), this document may need to be revised.

Keywords

IMT-2020, Fixed Network, Wireless Network, Network Performance, Quality of Service, Quality of Experience, QoS parameters, QoS classes

1 Scope

This document introduces the high-level end-to-end QoS/QoE requirements of IMT-2020. Basic terminologies and concepts are defined and attributes are developed independent of the underlying technology. With a view to facilitating future enhancements, the main body of the document describes the topic in general. It is noted that the descriptions of QoS in the document are designed to identify the customer's QoS/QoE requirements for various use cases and applications in IMT-2020, and the QoS requirements for service providers to offer could be derived accordingly. Consequently, future studies will be able to determine network technologies and architectures that can efficiently achieve the desired QoS/QoE requirements in IMT-2020.

The objective of this document is to describe the following:

- Overview of expected study areas, categorization, and definition of terminologies
- Gap analysis of activities of ITU and other organizations relevant to QoS/QoE of IMT-2020
- Reference models and performance parameters
- General guidance of performance parameters, QoS classes and their allocation
- Issues and items that need to be addressed by standards supporting IMT-2020

2 References

The following ITU-T Recommendations and other references contain provisions which through reference in this text, constitute provisions of this Recommendation. At the time of

publication, the editions indicated were valid. All Recommendations and other references are subject to revision; users of this Recommendation are therefore encouraged to investigate the possibility of applying the most recent edition of the Recommendations and other references listed below. A list of the currently valid ITU-T Recommendations is regularly published.

The reference to a document within this Recommendation does not give it, as a stand-alone document, the status of a Recommendation.

[ITU-T Y.1540] Recommendation (2011) - Internet protocol data communication service - IP packet transfer and availability performance parameter - http://www.itu.int/rec/T-REC-Y.1540-201103-I/en;

[ITU-T Y.1541] Recommendation (2011) - Network performance objectives for IP-based services - http://www.itu.int/rec/T-REC-Y.1541-201112-I/en;

[ITU-T Y.1542] Recommendation (2010) - Framework for achieving end-to-end IP performance objectives - http://www.itu.int/rec/T-REC-Y.1542-201006-I/en;

[ITU-T G.1000] Recommendation (2001) - Communications Quality of Service: A framework and definitions - http://www.itu.int/rec/T-REC-G.1000-200111-I/en

[ITU-T G.1010] Recommendation (2001) - End-user multimedia QoS categories - http://www.itu.int/rec/T-REC-G.1010-200111-I/en;

[3GPP TS 23.107] Technical Specification (2014) - Quality of Service (QoS) concept and architecture - http://www.3gpp.org/ftp/specs/archive/23_series/23.107/

[3GPP TS 23.401] Technical Specification (2015) - General Packet Radio Service (GPRS) enhancements for Evolved Universal Terrestrial Radio Access Network (E-UTRAN) access - http://www.3gpp.org/ftp/specs/archive/23_series/23.401/

3 Terms and Definitions

For further study

4 Abbreviations and acronyms

This deliverables defines the following abbreviations:

GBR	Guaranteed Bit Rate
GW	Gateway
IMT	International Mobile Telecommunications
IP	Internet Protocol
IPDV	IP packet Delay Variation
IPER	IP packet Error Rate
IPLR	IP packet Error Ratio
IPTD	IP packet Transfer Delay
MBR	Maximum guaranteed Bit Rate

NP	Network Performance
O&M	Operation and Maintenance
QCI	QoS Class Identifier
QoE	Quality of Experience
QoS	Quality of Service

5 Review of perspectives and standards on IMT-2020

5.1 Survey of Whitepapers

The first step in establishing a common QoS framework would be to identify the current status of QoS-related views and descriptions in relevant studies and white papers. This section therefore provides a survey on IMT-2020 QoS-related studies and white papers of the following organizations:

- International Organizations: NGMN, GSMA
- Regional Organizations: Horizon 2020, NetWorld2020, RAS Future Networks Cluster, 4G Americas
- Local Organizations: ARIB, Future Mobile Communication Forum of China, IMT-2020 Promotion Group, Huawei, ZTE, Nokia, Qualcomm, Ericsson, Samsung, NTT DoCoMo, Datang Telecom Technology & Industry Group

Type	Name	Views
International	NGMN	- Definition: No formal definition - General Requirement Description: Ability to enhance user experience and to provide differentiated and/or guaranteed QoS. - Parameters: User experienced data rate, latency, mobility, connection density, traffic density, coverage, signalling efficiency - Performance Objectives: 1Gb/s, 1~10ms, 500km/hr, 250k/km^2, 100Gbps/km^2, none, none
	GSMA	- Definition: No formal definition - General Requirement Description: Ability to maintain customer experience at peak data and to enhance quality of data roaming. - Parameters: Data rate, latency, bandwidth per unit area, number of connections, perception of availability, perception of coverage - Performance Objectives: 1-10Gbps, <1ms, 1000x LTE, 10-100x LTE, 99.999%, 100%
Regional	Horizon 2020	- Definition: No formal definition - General Requirement Description: Ability to provide differentiated and/or guaranteed QoS. - Parameters: Throughput, handover reliability, call drop rate, data access and discovery, signalling and traffic overhead - Performance Objectives: None
	NetWorld2020	- Definition (QoE): The degree of delight or annoyance of the user of an application or service. - General Requirement Description: Ability to optimize user QoE experience and to provide differentiated and/or guaranteed QoS.

		- Parameters: Bandwidth, delay, jitter - Performance Objectives: None
	RAS Future Networks Cluster	- Definition: No formal definition - General Requirement Description: Ability to provide differentiated and/or guaranteed QoS. - Parameters: Data rates, delay - Performance Objectives: 1000x LTE (100Gb/s per site), none
	4G Americas	- Definition: No formal definition - General Requirement Description: Ability to deliver the best QoS - Parameters: Data rate, latency, mobility, capacity, coverage - Performance Objectives: 100x, 5x~10x reduction, none, none, none
Local	ARIB	- Definition: No formal definition - General Requirement Description: Ability to maximize user perception with limited network resources. - Parameters: Peak data rate, capacity density, number of connected devices/cell, latency, mobility, reliability - Performance Objectives: >10Gbps, 1,000x IMT-Advanced, >10,000/cell, <1ms, >500km/hr, 90%~99.999%
	Future Mobile Communication Forum of China	- Definition: No formal definition - General Requirement Description: Ability to provide consistent user QoE. - Parameters: Peak data rate, guaranteed user data rate, connection density, traffic density, latency, mobility - Performance Objectives: >10Gbps, >100Mbps, >1M/km^2, >10Tbps/km^2, <1ms (radio) <10ms (E2E), up to 500km/hr
	IMT-2020 Promotion Group	- Definition: No formal definition - General Requirement Description: Ability to provide gigabit user experienced data rate and to satisfy QoS requirements of different application scenarios - Parameters: User experienced data rate, peak data rate, traffic density, connection density, latency, reliability - Performance Objectives: 100Mbps~1Gbps, >10Gbps, .10Tbps/km^2, 1M/km^2, <1ms, 100%
	Huawei	- Definition: No formal definition - General Requirement Description: Ability to resolve complex and different performance requirements by wide range of mobile services. - Parameters: Latency, simultaneous connections, data rate, switching time - Performance Objectives: <1ms, hundreds of billions of machines, 1Gb/s~10Gb/s, <10ms
	ZTE	- Definition: No formal definition - General Requirement Description: Ability to provide consistent on-demand access to services and to provide differentiated QoS. - Parameters: Capacity, peak data rate, latency, accuracy - Performance Objectives: None
	Nokia	- Definition: No formal definition - General Requirement Description: Ability to provide a virtual zero latency gigabit experience - Parameters: Capacity, latency, user data rates, coverage

		- Performance Objectives: 10,000x than LTE, <1ms, >10Gb/s, none
	Qualcomm	- Definition: No formal definition - General Requirement Description: Ability to enhance user experience and to provide differentiated and/or guaranteed QoS. - Parameters: Latency, reliability, coverage - Performance Objectives: None
	Ericsson	- Definition: No formal definition - General Requirement Description: Ability to deliver high-quality connectivity with very high availability. - Parameters: Capacity, data rate, latency, reliability, availability - Performance Objectives: None
	Samsung	- Definition: No formal definition - General Requirement Description: Ability to deliver uniform end-to-end experience regardless of user-location. - Parameters: Latency, peak data rate, cell edge data rate, mobility, simultaneous connections - Performance Objectives: <1ms, >10Gbps, >1Gbps, >500km/hr, >1M/km^2
	NTT DoCoMo	- Definition: No formal definition - General Requirement Description: Ability to provide more uniform user QoE than LTE - Parameters: Capacity, RAN latency, user throughput - Performance Objectives: 1000x compared to LTE, <1ms, 1Gbps everywhere
	Datang Telecom	- Definition: No formal definition - General Requirement Description: Ability to provide reliable performance. - Parameters: Peak data rate, latency, user throughput, throughput/km^2, connections/km^2, mobility - Performance Objectives: 10Gbps, <1ms, >10Mbps, >100Gbps/km^2, >1M/km^2, 1,200 km/hr

While most of the organizations agree that IMT-2020 network should provide both consistent user experience and differentiated/guaranteed QoS, the organizations differ in the specific aspects of network to be managed and therefore have different definitions of QoS. For example, some emphasize user experienced data rate and suggest monitoring of the parameter along with peak data rate. Nevertheless, some parameters (e.g., latency and mobility) are commonly proposed by the organizations and they may be able to provide a common ground for end-to-end QoS framework.

An important issue is that the views on IMT-2020 mostly focus on the Quality of Experience/Service at RAN (Radio Access Network), and seldom do organizations take into account end-to-end QoS (including RAN & Fixed Network). It is true that organizations such as Samsung and NGMN have acknowledged the importance of end-to-end latency in 5G network. However, most of the organizations suggest monitoring of user plane latency instead, and it is difficult to provide consistent QoS in this perspective.

5.2 Gap Analysis of Standards on IMT-2020

As noted in the previous section, current IMT-2020 QoS-related studies and views differ and are usually limited to RAN (Radio Access Network). While the need for standardization of single common end-to-end QoS is evident, the existing standards must be carefully studied in advance to identify topics for in-depth study. In this context, this section aims to provide a gap analysis20 of representative QoS standards of 3GPP and ITU-T.

Before proceeding to gap analysis, some major terms must be defined to avoid confusion. 3GPP simply defines a concept of bearer, a link between two end-points defined by a certain set of characteristics, to implement QoS. ITU-T, on the other hand, distinguishes NP (Network Performance) from QoS to translate user demands to network operational attributes. While QoS is defined to be "collective effect of service performances which determine the degree of satisfaction of a user of the service," NP is defined independently of terminal performance and user actions to provide information for system development, network planning, and O&M (operation and maintenance). In other words, QoS views the "quality" of network in the user's perspective and the other two (bearer and NP) views the same concept in network providers' perspective. This document will follow the definitions as provided by 3GPP and ITU-T.

5.2.1 Definition of "End-to-End" in different standards

3GPP's concept of "end-to-end" comprehensively covers the whole network from a user's device to another user's device. However, its UMTS bearer concept is limited to an interval starting from user's device to PDN gateway (a gateway in wireless core network) for the sake of practicality (i.e., a network operator can influence only its network and its radio interface). (3GPP TS 23.107, TS 23.401 Rel.12)

ITU-T, on the other hand, attempts to identify network QoS from end user to end user by defining UNI (User Network Interface) to UNI objectives in Y.1541. The concept, however, is usually applied to wireline IP-based services without any specific discretion on technologies of lower layers.

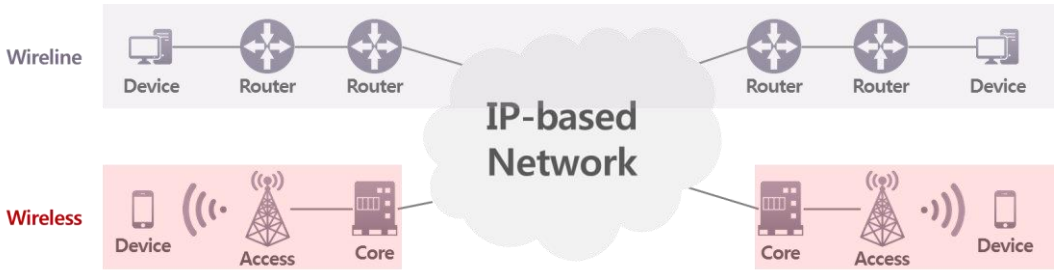


Figure 1. The scope of 3GPP (Red) and ITU-T Standards (Purple)

5.2.2 Layered Approach for QoS Management

The all-IP nature of IMT-2020 network allows data to be transported without connection on IP layer. With a given pair of source and destination IP address, IP packets are transferred from an end to another. The performance of an IP service, however, also depends on the performance of other layers (both upper and lower layers) and it is important to identify/acknowledge the relationship between performances of different layers.

---

20 IMT-2020 network is predicted to be all-IP environment and it will be sufficient to compare standards covering QoS of IP networks as a preliminary step

ITU-T's Y.1540 standard provides a layered model of performance of IP service to illustrate the point aforementioned. The lower layers do not have end-to-end significance (i.e., it transfers packet from a point to another) but the type of technology employed (e.g., Ethernet-based leased lines) may affect the performance. Higher layers may also affect performance. 3GPP's bearer acknowledges the effect of various layers on IP services, but defines the bearer on layer 1 and 2 for the use of higher layers (3GPP TS 23.107 & 23.401). Nevertheless, both acknowledge that the framework must take into account the impact from performance of layer 1 and 2 in both wireline and wireless media.

5.2.3 QoS Classification

Not only is the scope of the standards different, but also classification of QoS. 3GPP (TS 23.107, 23.203, 23.401) classifies QoS bearer using QCI (QoS class identifier) and ARP (Allocation and Retention Priority). QCIs provide the thresholds of basic parameters such as packet delay and packet error loss rate, while ARP provides the basis for establishing different bearers in the face of resource limitations (i.e., it prioritizes the allocation and retention of bearers not packets). The bearers are provided as either GBR (bearer with the minimum guaranteed bit rate per bearer) or MBR (the maximum guaranteed bit rate per EPS bearer). In other words, some applications (mission critical applications such as conversational voice and conversational video) are to be provided via GBR bearer and other applications are to be provided via MBR bearer.

ITU-T, on the other hand, specifies six classes (one of whose parameters are unspecified) of QoS to differentiate and guarantee traffic quality. While the details of classification is different, 3GPP's basic philosophy of differentiating mission-critical class traffic from default best-effort traffic is also inherent in ITU-T's QoS classification.

QCI	Type	Packet Delay (ms)	Packet Error Loss Rate	Example Services
1	GBR	100	10^{-2}	Conversational Voice
2		150	10^{-3}	Conversational Video (Live Streaming)
3		50	10^{-3}	Real Time Gaming
4		300	10^{-6}	Non-Conversational Video (Buffered Streaming)
65		75	10^{-2}	Mission Critical user plane Push To Talk voice
66		100	10^{-2}	Non-Mission-Critical user plane Push To Talk Voice
5	MBR	100	10^{-6}	IMS Signalling
6		300	10^{-6}	Video (Buffered Streaming), TCP-based services
7		100	10^{-3}	Voice, Video (Live Streaming), Interactive Gaming
8		300	10^{-6}	Video (Buffered Streaming), TCP-based services
9		300	10^{-6}	Video (Buffered Streaming), TCP-based services
69		60	10^{-6}	Mission Critical delay sensitive signalling
70		200	10^{-6}	Mission Critical Data

Table 1. 3GPP's QoS Classification (QCI only)

QoS Class	IPTD	IPDV	IPLR	IPER
0	100ms	50ms	10^{-3}	10^{-4}
1	400ms	50ms	10^{-3}	10^{-4}
2	100ms	-	10^{-3}	10^{-4}
3	400ms	-	10^{-3}	10^{-4}
4	1s	-	10^{-3}	10^{-4}
5 (Unspecified)	-	-	-	-

Table 2. ITU-T's QoS Classification (QoS Class)

5.3 Rationale for a common QoS framework for IMT-2020

The differing and RAN-centric view of IMT-2020 QoS calls for a systematic and integrated approach to establish a common framework for end-to-end QoS. Since the content of the existing standards are not sufficient, this document aims to provide a common single end-to-end standard.

6 Network performance, Quality of Service and Quality of Experience

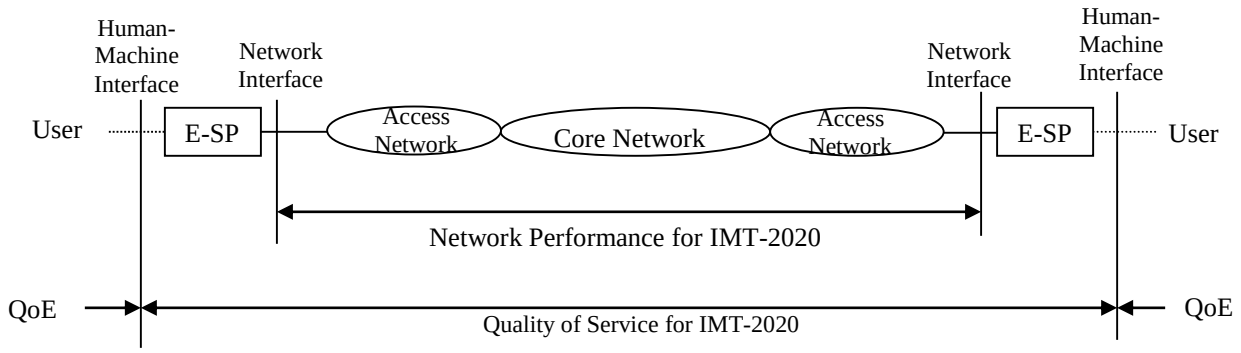
QoS is defined in Recommendation E.800 as follows: "Collective effect of service performance which determine the degree of satisfaction of a user of the service".

The definition of QoS in Recommendation E.800 is comprehensive and encompasses many areas of work, including subjective user satisfaction. However, within this document the aspects of QoS that are covered are restricted to the identification of parameters that can be directly observed and measured at the point at which the service is accessed by the user.

Recommendation I.350 defines Network Performance as follows: "NP is measured in terms of parameters which are meaningful to the network provider and are used for the purpose of system design, configuration, operation and maintenance. NP is defined independently of terminal performance and user actions."

QoE is defined as the overall acceptability of an application or service, as perceived subjectively by the end-user. Quality of Experience includes the complete end-to-end system effects. Overall acceptability should be influenced by user expectations and context.

Figure 1 illustrates how the concepts of QoS, NP and QoE can be applied in IMT-2020.



- ESP: End-Service Platform (i.e., Mobile/smart phone, data server, appliances, TV, etc.)

Figure 2. General reference configuration for QoS, NP and QoE

Editor's Note 1: User-to-user communication is the basic consideration in this figure. In order to cover the IMT-2020 in broader perspective, the connectivity of IMT-2020 should facilitate Machine-to-machine/Device-to-device interfaces as well. This is for further study.

Editor's Note 2: The above definitions and configuration are derived from existing general telecommunication terminologies. Based on the study result of new architecture for IMT-2020, those definitions and configuration (i.e., the coverage or location of wireless interface) may be changed. This is for further study.

QoS, NP and QoE are related concepts with different focus and scope.

QoS provides a valuable framework for network provider, but it is not necessarily usable in specifying performance requirements for particular network technologies (i.e. ATM, IP, MPLS, etc.). Similarly, NP ultimately determines the (user observed) QoS, but it does not necessarily describe that quality in a way that is meaningful to users.

QoE is subjective in nature, i.e. depend upon user actions and subjective opinions.

The definition of QoS, NP and QoE should make mapping clear in cases where there is not a simple one-to-one relationship among them.

Table 1 shows some of the characteristics which distinguish QoS, NP and QoE.

Table 3. Distinction between quality of experience, quality of service and network performance

Quality of Experience	**Quality of Service**	**Network Performance**
User oriented		Provider oriented
User behaviour attribute	Service attribute	Connection/Flow element attribute
Focus on user-expected effects	Focus on user-observable effects	Focus on planning, development (design), operations and maintenance
User subject	Between (at) service access points	End-to-end or network elements capabilities

The separation of QoS, NP and QoE indicates that development of corresponding parameters should take into account the following general points:

- the definition of QoS parameters should be clearly based on events and states observable at service access points and independent of the network processes and events which support the service;
- the definition of NP parameters should be clearly based on events and states observable at network element boundaries, e.g. protocol specific interface;
- the definition of QoE parameters is for further study.

7 Top down perspective of QoS

Applying QoS in the field usually takes four steps, which is well illustrated and recommended by Figure 1 of ITU-T Recommendation G.1000 as follows:

① Customer's QoS requirements: This is a perspective focusing on the resulting end-to-end service quality. The customer is not concerned with how a particular service provided or with any aspects of the telecommunication network's internal design and operation.

② Service provider's offerings of QoS (or planned/targeted QoS): This is the level of quality expected to be offered to the customer by the service provider. It is expressed by values assigned to QoS parameters and examples include SLA (Service Level Agreements).

③ QoS achieved or delivered: This is the quality actually achieved and delivered to the customer. This is expressed by values assigned to parameters so that it can be compared to the offered QoS.

④ Customer survey ratings of QoS: This is the level of quality that customers believe they have experienced. It is usually expressed in terms of degrees of satisfaction rather than technical terms. These survey ratings are then reflected in the customer's QoS requirements and other viewpoints will be revised accordingly.

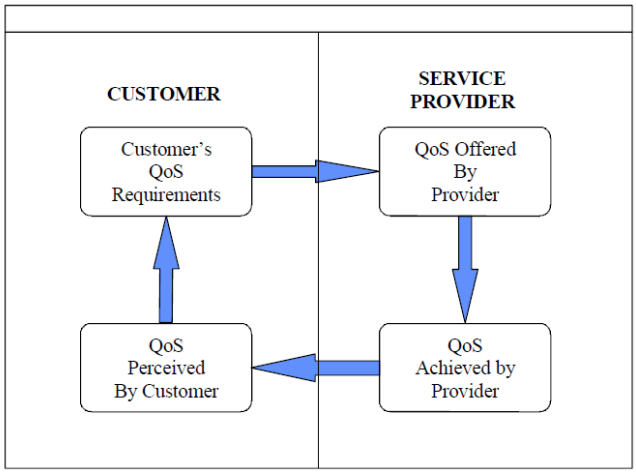


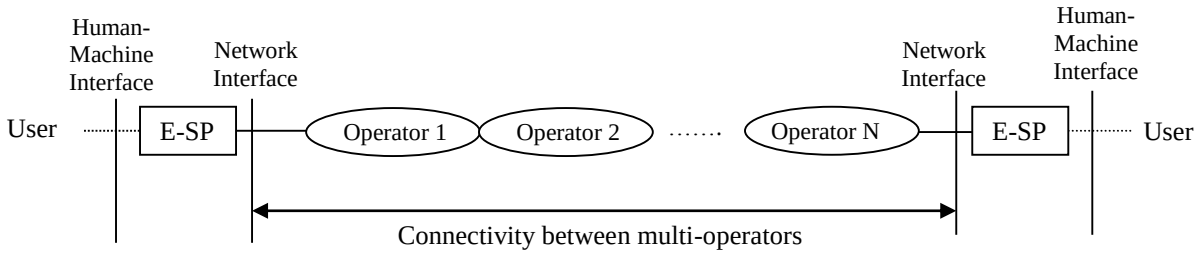
Figure 3/G.1000(Figure 2). The four viewpoints of QoS

Among the four viewpoints, the scope of this focus group will mainly focus on the first perspective of top down approach: defining customer's QoE requirement for IMT-2020. While the second to fourth steps are essential for building the QoS framework, they are not necessarily included in the scope of this focus group. This is because the three steps are dependent on protocols and/or technologies used in the telecommunication network, and these issues may be resolved by further study in corresponding study groups.

8 Reference model of Connectivity

Variations in connectivity configurations and diversity of use cases make development of a single universal figure for end-to-end IMT-2020 connectivity impossible. Since prospective IMT-2020 applications and legacy telecommunication services have to cover short-range communication and inter-continental communication, various connectivity configurations should be considered. For example, the longest connectivity will have length of 27,500km (half of the equatorial circumference of the Earth, ITU-T Recommendation G.801) while V2V (Vehicle-to-vehicle) communication can take place in the connectivity of less than10m.

The following figure presents a general reference configuration for the connectivity of IMT-2020. The connectivity may be delivered via both wireless and wireline media.



- E-SP: End-Service Platform (i.e., Mobile/smart phone, data server, appliances, TV, etc.)

Figure 4. General reference configuration for the connectivity of IMT-2020

Editor's Note: The above definitions and configuration are derived from existing general telecommunication terminologies. Based on the study result of new architecture for IMT-2020, those definitions and configuration (i.e., the coverage or location of wireless interface) may be changed.

Various connectivity configurations that can be considered are illustrated in figure 2 (listed in descending order of physical distance):

① International (including inter-continental) inter-operator communication

② Intra-national inter-operator communication

③ Intra-operator communication

④ Device-to-Network communication

⑤ D2D (device-to-device) communication

* Note: configurations #1 ~ #3 each consist of wireless-wireline, wireline-wireline, and wireless-wireless communications, but figure 5 shows only wireless-wireline communication for simplicity

While the first three connectivity cases are already utilized by the legacy telecom service, IMT-2020 specific connectivity (i.e., case #4&5) may not fully supported by legacy connectivity and this calls for a new approach. Traditional SDOs (ITU, 3GPP, etc.) have already defined QoS requirements of the first three connectivity cases excellently and relatively minor modifications will be necessary for them to be applied to IMT-2020. However, IMT-2020 specific connectivity (such as case #4&5) is not covered by existing standards and a new QoS framework to address the case is necessary.

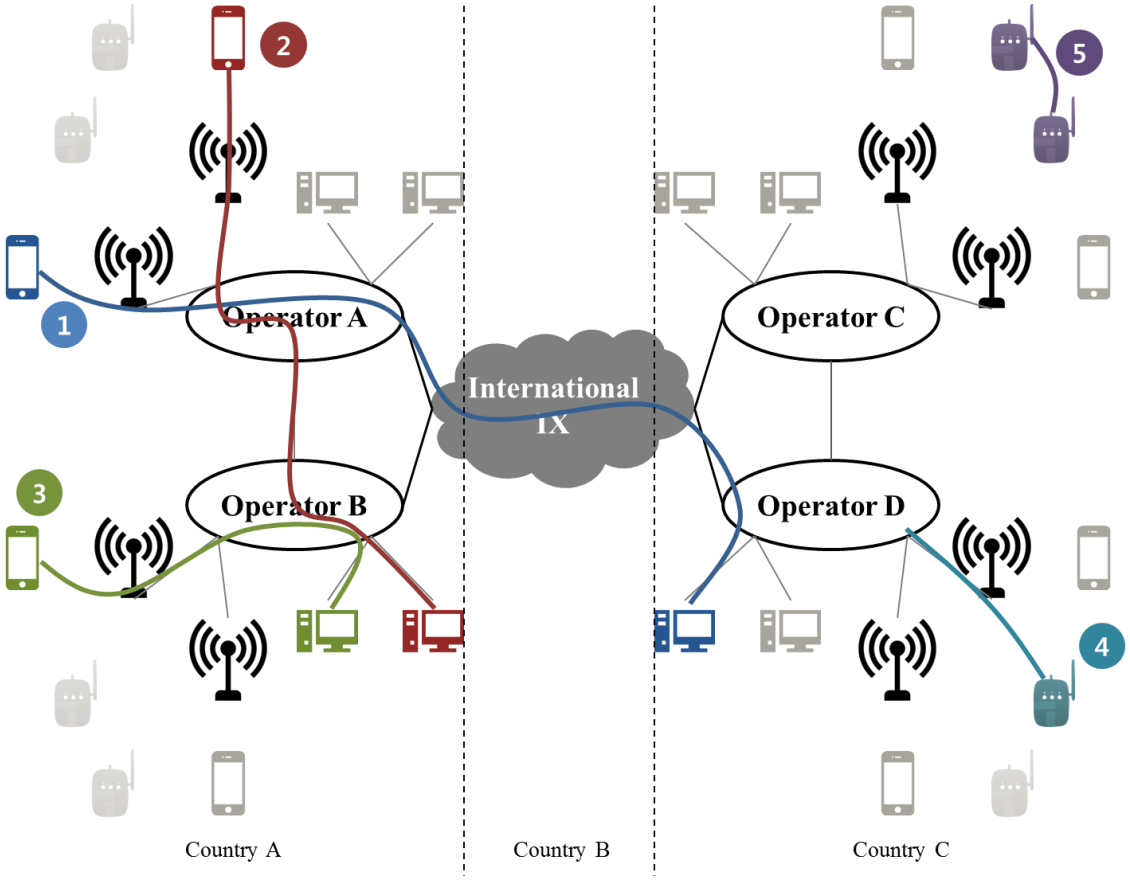


Figure 5. Various configurations of connectivity

The diversity of configurations suggests that high-level end-to-end QoE requirement for IMT-2020 needs to consider the distance and the complexity (i.e., number of telecommunication systems) of the configuration in concern. Therefore, the standard should define allocation rules based on longest reference case of connectivity and requirements for specific use cases based on shorter reference case of connectivity.

9 Layered model of performance

For operators to realize the customer QoE requirements in IMT-2020 network, QoS engineering is required for different network technologies and protocols.

While this document focuses on the concept of connectivity, which is independent of underlying technologies/protocols, it is nevertheless necessary to consider their effects in different layers. As an example, the delay perceived by user may consist of delays from the processing in application layer, the transfer of packets in IP layer, the overhead of interworking different protocols in the data link layer, and the delays in traveling from one source to another via a physical medium. For example, a VoIP call may have total accumulated delay of 100ms that consists of the following delays: 40ms processing delay in user handset, 50ms delay due to routing and 10ms propagation delay over fibre optic links.

Also, having different media/protocol in the same layer affects the performance. For example, the distance and the complexity of connectivity must be taken into account to define delay-related

requirements. If end-to-end connectivity with delay of less than 10ms is necessary, the connectivity should be designed within 1,600km of fibre optic links (because fibre optic link will have a propagation delay of 6.25μs/km as defined in ITU-T Recommendation I.356). Then the processing and queueing delays in a network system (i.e., router, switch, etc.) should be considered. If a network system issues 2ms delay for internal processing and queueing and 3 network systems are required in the connectivity, the distance of the link must be reduced accordingly into 640 km. If servers for specific services are required in the connectivity, these factors should be taken into account as well. Furthermore, if wireless links are employed in this connectivity, the connectivity must be engineered to suit the protocols and the technologies of wireless links.

In other words, realizing customer's QoE requirements must be based on the structure of technologies and protocols in different layers. This is why understanding of different layers of networks is indispensable.

The figure below describes the general layered model that could serve as a reference for IMT-2020. Detailed layered models for different use cases and scenarios are for further study.

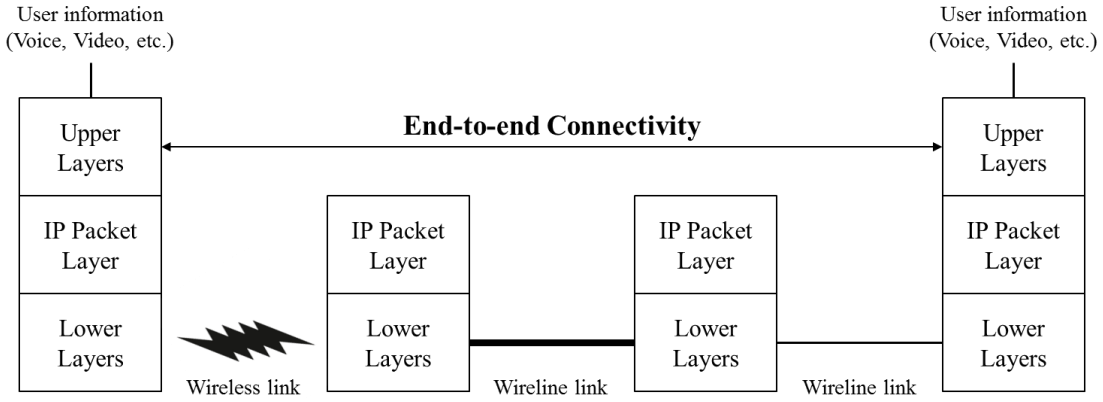


Figure 6. Layered model of performance for IP service in IMT-2020

*** Examples of lower layers are,**
Wireline: MPLS, Ethernet, Optic Fibre, etc.
Wireless: PDCP, RLC, Air interface, etc.

10 QoE parameters

Various IMT-2020 standards differ in terms of parameters employed to evaluate QoS. Since QoS parameters are the basic building block of QoS framework, it is necessary to define QoS parameters for IMT-2020.

In this context, this section proposes two methodologies for defining QoS parameters of IMT-2020 to build the fundamentals of common end-to-end QoS framework for IMT-2020.

1) To select relevant QoS parameters from existing standards and integrate the selections in one coherent framework

2) To newly define QoS parameters from zero base that are suitable for IMT-2020 network

10.1 Selecting QoS Parameters from different standards

To select relevant QoS parameters from existing standards, it is necessary to identify the QoS parameters employed by existing standards. This section will therefore review the parameters of wireline standard from ITU-T and wireless standard from 3GPP.

10.1.1 ITU-T Y.1540, Y.1541

ITU-T recommendation Y.1540 defines different QoS parameters and Y.1541 selects four major parameters to classify QoS classes: IPTD (IP Packet Transfer Delay), IPLR (IP Packet Loss Ratio), IPER (IP Packet Error Ratio) and IPDV (IP Packet Delay Variation).

IPTD is the time between the occurrences of two corresponding IP packet reference events. IPTD in this context denotes mean IP packet transfer delay performance parameter, which is the arithmetic average of IP packet transfer delays for the population of interest.

IPLR is the ratio of total lost IP packet outcomes to total transmitted IP packets in a population of interest. In other words, it describes how many packets were lost relative to number of packets transmitted.

IPER is the ratio of total errored IP packet outcomes to the total of successful IP packet transfer outcomes plus errored IP outcomes in a population of interest. In other words, it describes how many packets faced error out of all packets that were transferred.

IPDV is the packet delay variation in the two-point interval experienced relative to the reference IP packet transfer delay (it is the absolute IP packet transfer delay experienced by a selected IP packet between the two points).

Please note that the parameters on end-to-end interval are derived from their counterparts in segments of the network and the method is described in Y.1541.

10.1.2 3GPP TS 23.107, 3GPP TS 23.203, 3GPP TS 23.401

3GPP, as noted in previous input documents, defines QoS with the concept of bearer. Bearer is a link between two end-points defined by a certain set of characteristics (on layers 1 to 2/2.5). The QoS is then controlled by two major parameters: QCI (QoS Class Identifier) and ARP (Allocation and Retention Priority).

Holistically, ARP is the standard for prioritizing bearer establishment and QCI is the standard for prioritizing packet transfer. ARP will consist of the priority level and parameters related to “pre-emption” of resources.

QCI, on the other hand, provides detailed QoS information for types of bearers. Firstly, bearers are defined either as GBR (Guaranteed Bit Rate) or MBR (Maximum Bit Rate) where the former serves to guarantee the throughput. Secondly, bearers possess thresholds of parameters such as packet delay and packet error loss rates to ensure QoS.

Packet delay is an upper bound for the time that a packet may be delayed between the UE (user equipment) and PCEF (Policy and Charging Enforcement Function). It is, in other words, a maximum delay with a confidence level of 98 percent. Packet error loss rate is an upper bound for the rate of SDUs (Service Data Unit; e.g., IP packets) that have been processed by the sender of a link layer protocol but that are not successfully delivered by the corresponding receiver to the upper layer. In other words, it defines an upper bound for a rate of non-congestion related packet losses. (3GPP TS 23.203)

10.2 Approach for defining new parameters

This section proposes to utilize the framework by ITU-T Recommendation I.350 in defining new parameters.

Recommendation I.350 defines 3x3 matrix for QoS parameters with each row representing one of the three basic and distinct communication functions (access [21], user information transfer [22] and disengagement [23]) and each column representing one of the three mutually exclusive outcomes possible when a function is attempted (speed[24], accuracy[25] and dependability[26]). Table 2 shows the defined 3x3 matrix for QoS parameters. The parameters are then mapped onto the matrix with outage thresholds defined.

Table 4. 3 x 3 matrix approach for QoS parameters

	Speed	Accuracy	Dependability
Access			
User information transfer			
Disengagement			

The 3x3 performance matrix may be extended to address QoS and NP of IMT-2020. The performance criterion might be supplemented with criteria such as ease-of-use. The communications functions might be supplemented with information storage, information translation, and brokerage functions. Such addition of functions and criteria is possible, but the functions should be quantifiable.

Identification of each affordable parameter is for further study.

Editor's Note #1: Figure1/G.1000 shows the extended matrix between criteria and function to facilitate identification of communications QoS criteria.

Editor's Note #2: The parameters should take into account the characteristics of IMT-2020 specific applications such as remote surgical operation, autonomous driving and virtual reality. The details are for further study

Editor's Note #3: In addition, for Device-to-Device and Device-to-Network cases with very low delay (1ms), definition of Measurement Point and Monitoring Methodology is critical. The details are for further study.

11 QoS classes and their performance objectives

QoE requirements of the applications/services for IMT-2020 can be identified by an extremely wide range from best effort to very stringent level.

---

21 Issuance of an access request signal or its implied equivalent at the interface between a user and the communication network

22 Begins on completion of the access function and ends when the "disengagement request" is issued. It includes all formatting, transmission, storage, error control and media conversion operations performed on the user information during this period

23 Issuance of a disengagement request signal

24 Describes the time interval that is used to perform the function or the rate at which the function is performed

25 Describes the degree of correctness with which the function is performed

26 Describes the degree of certainty (or surety) with which the function is performed regardless of speed or accuracy

These various levels should be classified as follows;

1) Based on end-to-end user expectation of impairments and is therefore not dependent on any specific technology (network as well as application) for its validity. But the classification should be easily applied to network technologies for the purpose of implementation and operation.
2) Shows how the performance parameters (delay, delay variation, loss, etc.) and their objectives can be grouped appropriately, with implying that one class may "better" than another.

Specific QoE classes and their performance objectives for IMT-2020 are for further study.

Editor's Note: This is an important topic and we are more than welcome to have suggestions

12 Allocation guidance

IMT-2020 will consist of various use cases that will require different connectivity configurations. This means that the QoS framework has to take into account various cases in order to ensure QoS of connectivity in IMT-2020.

One of the aspects that deserve special treatment is QoS budget allocation (also known as impairment allocation). Although end-to-end QoE requirement will be defined, the implementation will be different for various networks with different circumstances. This calls for an in-depth study on budget allocation approaches.

In this context, this section outlines different methodologies for QoS budget allocation and identifies additional consideration to be considered in the IMT-2020 QoS framework.

12.1 Methods for QoS Budget Allocation

This section will outline QoS budget allocation methods as described in ITU-T Recommendation Y.1542 and identify the need for a new approach for IMT-2020 specific connectivity cases. If details are necessary for the first four subsections, please consult Y.1542.

12.1.1 Static Approach

This approach divides the UNI-to-UNI path into a fixed number of segments and budgets the impairments such that the total objective is met in principle. It requires that individual segments have knowledge of the distance and traffic characteristics between the edges of their domains, as these properties of the segment affect the resulting allocations.

An important aspect of the static allocation is its dependence on the number of providers, as the allocation has to be done accordingly. This can result in undershooting or overshooting the objective because any actual path may traverse a different number of network segments from what was assumed to be the case in the allocation scheme.

12.1.2 Pseudo-static Approach

In this approach, each provider would have knowledge of how many providers are present in the traffic path and allocate among each other without wasting part of the impairment budget. Service providers may reallocate their impairment target among the segments under their control.

12.1.3 Signalled Approach

In this approach, providers will use signals to communicate and determine impairment budgets. In this approach, the use of resource management and signalling for the purposes of impairment apportionment is assumed. This section will consider only two kinds of signalled approach for simplicity.

The first type of signalled approach is negotiated allocation approach. In this approach, networks negotiate with one another in allocating impairment budget. Starting with initial segment impairment targets, based possibly upon the static and pseudo-static allocations, the networks may negotiate for any "impairment budget" excesses, and to advertise to multiple interested parties if they can provide a network service that is within their collective impairment budget. If it is not possible to do so, the network can ask the previous network (or the user) whether more impairment budget can be allocated such that the delivery path can be determined.

The second type of signalled approach is ranged allocation approach. In this approach, the range between the minimum and maximum of the allocated impairment budget for every segment along the data path is negotiated and calculated out by the use of resource management and signalling among the segments. Any value within each segment impairment budget range, when added with those of other segments, can meet the total impairment budget target for the whole data path. Thus, every segment itself can choose an appropriate value within its allocated budget range under the consideration of optimizing its resource utilization.

12.1.4 Impairment accumulation Approach

This approach is where possible performance levels offered by each provider are used to calculate the estimate of UNI-UNI performance and lead to decisions on path/QoS. The mechanism starts with a requesting provider determining a path that packets will follow and requesting each provider for the performance level that each will commit to. With the offers, the requesting provider will estimate the overall performance level and compare it with the UNI-to-UNI QoS class/objectives. If the path does not meet the requested objectives, the provider could take one of the following actions

1. Path negotiation: An alternative path might be sought (repeating the request and comparison process for the new path)
2. User negotiation: An alternative service class or relaxed objectives could be offered.

12.1.5 Need for a new approach

As noted in section 7, there are IMT-2020 specific connectivity cases (cases #4 and #5) that deserve special treatment. Device-to-network communication is different from conventional communication in aspects such as frequency of communication (periodic) and type of traffic generated (usually more signalling traffic than data). Device-to-device communication also is distinctly different from conventional communication because the distance will be much shorter and the configuration will be simpler (with smaller number of nodes). The differences show that in-depth study is necessary to develop QoS budget allocation for these connectivity configurations.

Editor's Note: This is an important topic and we are more than welcome to have suggestions

13 Examples of end-to-end connectivity over wireless and wireline networks

Editor's note: This clause was originally indicated as an appendix in the draft output from the end-to-end QoS group.

13.1 International Voice Call

When devising SLA (service level agreements) of international voice calls for enterprise customers, international standards provide legal ground. The issue in this business is that the recommendation of 150ms delay (or latency) in one way (customer's QoS requirements from mouth to ear connectivity) is enforced for an operator in concern for natural customer experience, and end user will suffer from longer delay than is acceptable.

As indicated in Input Document I-011 submitted to the first Focus Group meeting, there are, however, two different relative standards for wireline and wireless networks respectively. ITU-T Recommendation Y.1541 specifies end-to-end delay to 100ms (excluding 50ms processing delays in both end terminals) for real-time voice application on QoS class 0 in wireline network as 3GPP TS 23.107 Rel.12 does so for conversational voice on QCI 1 in wireless network.

The resulting delay of a voice call is likely to exceed the threshold of 150ms. When domestic wireless operator and an international wireline network operator maintain their QoS level at 90ms each, both of them have fulfilled their legal responsibility of SLA, but the end user will experience QoS of 180ms delay, which is not tolerable and is worse than the customer's expectation (100ms).

Since IMT-2020 network should consider some cases where services are provided over wireless and wireline networks of different operators, the current disparate standards on wireless and wireline networks are not suitable for providing customer QoS requirements. Therefore, a new standard applicable on both wireless and wireline networks is necessary for IMT-2020.

13.2 Domestic Video Telephony

With the proliferation of smart devices and electronic devices, video telephony has become a common service for the public. Since smart devices are connected to the internet via WiFi and Cellular, and PC via wireline, integrated perspective (wireline & wireless) in domestic connectivity on QoS management is necessary for ensuring customer satisfaction of video telephony QoS.

Video Telephony as used here implies a full-duplex system, carrying both video and audio intended for use in a conversational environment. As such, the same delay requirements as for conversational voice will apply in principle (i.e. no echo and minimal effect on conversational dynamics) with the added requirement that the audio and video must be synchronised within certain limits to provide "lip-synch".

The quality of video must take into account the nature of human eye. Human eye is tolerant to some loss of information, so that some degree of packet loss is acceptable depending on the specific video coder and amount of error protection used. It is expected that the latest MPEG-4 video codecs will provide acceptable video quality with frame erasure rates up to about 1%. It should be noted that the QoS of video (in video telephony) may vary depending on compression ratio and application (i.e., real-time streaming, buffered streaming, real-time conversation, etc.), but the QoS threshold of voice should be 150ms as noted in 3.1.

Regarding the QoS of video telephony, Input Document I-011 submitted to the first Focus Group meeting describes the definition by current standards. ITU-T Recommendation Y.1541 specifies end-to-end delay to 100ms (excluding 50ms processing delays in both end terminals) and 10^{-3} Packet Loss Ratio for real-time video telephony application on QoS class 0 in wireline network

while 3GPP TS 23.107 Rel.12 does 150ms delay and 10^{-3} Packet Error Loss Rate for conversational video on QCI 2 in wireless network.

The resulting delay of a voice call is likely to exceed the threshold of 150ms. When domestic service operator maintain their QoS level at 90ms for wireline and 120ms for wireless each, both of the networks have fulfilled their standard-based operation, but the end user will experience QoS of 210ms delay, which is not tolerable and is worse than the customer's expectation (100ms).

Since IMT-2020 network should consider the case where a service provided over wireless and wireline networks within a single operator, it is not suitable to apply current disparate standards within the operator. This calls for a new standard applicable on both wireless and wireline networks for IMT-2020.

13.3 Telecommunication services for Emergency/Disaster Relief

In an emergency, available network resources are dramatically reduced and calls for a prioritization of communication requests. Some forms of communications, such as calls between disaster-related government agencies, have greater importance than others, or perform a more critical function than other forms of communication. Classifying priority classes and assigning relative priorities can help enhance efficient and timely use of network resources.

In this case, however, a single standard encompassing wireless and wireline networks is necessary for ensuring consistent customer QoS requirement and enhancing operational efficiency in IMT-2020.

In addition, some parts of the network (whether randomly located or concentrated) will be damaged and must be replaced by temporary measures during disaster. In this process, what used to be wireline network may be replaced by temporary wireless networks and the opposite may hold true. It is therefore essential to define the comprehensive (wireless & wireline) QoS requirements of disaster relief systems independent of the type of technologies/protocols across the network even during disasters.

It should also be noted that communications during disaster take place in various forms such as voice, SMS/MMS and video streaming. These forms of communications need to satisfy very stringent QoS requirements; 1) exact location information of casualties, 2) relief instructions from control tower, 3) high-definition real-time video of the disaster site for first aid, etc.

While cases similar to the ones presented in 3.1 and 3.2 will occur, communication may have to go through from a wireless network to wireline network in a single operator to another operator's network. The worst case scenario will be going from an operator A's wireless network to operator B's wireless network via each operator's wireline network. In this case, each operator and each network may maintain QoS level at 90ms, each fulfilling own legal responsibilities. Yet, the end users will perceive an overall delay of 360ms and it will be difficult to provide high-definition real-time forms of communications.

Since current standards employ different QoS parameters (e.g., 3GPP defines the type of traffic, packet delay and packet loss while ITU-T Y.1541 defines IP packet delay, IP packet delay variation, IP packet loss and IP packet error), communication during disaster will increase complexity of operation and may lead to confusion of QoS management.
In addition to a common QoS parameter, methods for QoS implementation (e.g., resource reservation and priority control) can also be matched. While cellular communications employ pre-emptive measures to guarantee QoS, wireline communications process priorities of individual packets. For consistent operation and management, it is recommended that a common QoS implementation method be applied over both wireline and wireless networks.

Appendix IV

Mobile front haul and back haul

Editor's Note: Appendix IV was produced during the FG-IMT 2020 focus group in order to investigate gaps in standardization related to IMT-2020. While the request from SG-13 was to deliver a report outlining standardization gaps, the consensus of the focus group was that the working documents produced and used during the focus group work contained useful information for future work and should be captured. Note, however, the focus group concentrated on producing accurate descriptions of the standardization gaps in the main body of this document; some minor errors may exist in the appendices. They are, however, the output of the focus group but are provided for information only.

Editor's Note: This appendix uses clause references in a form usually associated for normative text. This is maintained for this report to align with references made in the main body of this report. The original text was assuming and Appendix D, not the current Appendix IV

This document contains the baseline of Annex D going into the Beijing meeting. Comments discussed on the Oct. 14 teleconference have been addressed.

Annex D: Mobile front haul and back haul – Drivers, Challenges, and Solutions

1. Scope

This document contains material relevant to Mobile Front-haul and Mobile Back-haul.

2. References

[Cisco] Cisco Visual Networking Index (VNI), "Global Mobile Data Traffic Forcast Update," Feb. 2013, http://www.gsma.com/spectrum/wp-content/uploads/2013/03/Cisco_VNI-global-mobile-data-traffic-forecast-update.pdf

[Detnet] IETF Deterministic Networking, http://trac.tools.ietf.org/bof/trac/wiki/DetNet.

[Docomo] Y. Shimazu, H. Ohyane, T. Watanabe, T. Yajima and S. Suwa: "LTE Base Station Equipment Usable with W-CDMA System," NTT DOCOMO Technical Journal, Vol. 13, No. 1, pp.20-25, June 2011. (https://www.nttdocomo.co.jp/english/binary/pdf/corporate/technology/rd/technical_journal/bn/vol13_1/vol13_1_020en.pdf)

[ITU-R_F15] Report ITU-R M.2375-0, "Architecture and topology of IMT networks" Fig.15, p40, Oct.,2015

[ITU-R_F15] Report ITU-R M.2375-0, "Architecture and topology of IMT networks" Fig.16, p40, Oct.,2015

[MEF 22.1.1] MEF 22.1.1- Mobile Backhaul Implementation Agreement

[MIC] Ministry of Internal Affairs and Communications, "2011 WHITE PAPER on Information and Communications in Japan," Part 1, Section 1 page 1, http://www.soumu.go.jp/johotsusintokei/whitepaper/eng/WP2011/part1.pdf.[MSR] Mobile

Society Research Institute, NTT DOCOMO, INC, “Disaster resistant information society”, NTT Publishing Co., Ltd., 2013 (in Japanese).

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

[CPRI]; Common public radio interface (CPRI)-interfaces specification.

[TSN] IEEE Time-Sensitive Networking Task Group, http://www.ieee802.org/1/pages/tsn.html.

[TTC] TTC white paper “white paper on Future Networking issue”, Apr., 2015

http://www.ttc.or.jp/e/topics/20150413/download/

3 Terminology Definitions

3.1 Backhaul refers to the network paths connecting the Base Station sites and the Network Controller/ Gateway sites.

3.2 Fronthaul refers to the intra-base-station transport, in which a part of the BS function is separated to the remote antenna site. (Note that this definition is equivalent to the definition given in [MEF 22.1.1] for the current 4G technology.)

4 Abbreviations and acronyms

4G: 4th Generation (mobile system)
5G: 5th Generation (mobile system)

ADC: Analogue Digital Converter
a.k.a. : also known as
API: Application Programming Interface
AR: Augmented Reality

BBU: Base Band Unit
BDE: Base station Digital processing Equipment
BH: Backhaul
BRE: Base station Radio processing Equipment
BS: Base Station

C-RAN: Centralized Radio Access Network
CB: Coordinated Beam Forming
CoMP: Coordinated Multi-Point
CPRI: Common Public Radio Interface
CS: Coordinated Scheduling

DAC: Digital Analogue Converter
DBA: Dynamic Bandwidth Allocation
DPS: Dynamic Point Selection
DSP: Digital Signal Processing

E2E: End to End
EVM: Error Vector Magnitude

FH: Fronthaul

IEEE: The Institute of Electrical and Electronics Engineers, Inc.
IETF: Internet Engineering Task Force
IMT: International Mobile Telecommunication
IoT: Internet of the Things
IoE: Internet of Everything
IP: Internet Protocol
ITU-T: International Telecommunications Union

JR: Joint Reception
JT: Joint Transmission

LH: Long Haul
LRE: Low-power Radio Equipment

M2M: Machine to Machine
MAC: Media Access Control
MAN: Metropolitan Area Network
MBH: Mobile Backhaul
MEF Metro Ethernet Forum
MFH: Mobile Fronthaul
MIMO: Multi-Input Multi-Output
MPLS: Multi-Protocol Label Switching
MVNO: Mobile Virtual Network Operator

NFV: Network Function Virtualization
NGFI: Next Generation Fronthaul Interface
NGMN: Next Generation Mobile Network

ODN: Optical Distributed Network
OFDMA: Orthogonal Frequency-Division Multiple Access
OAM: Operation and Management
OPEX: Operating Expense
OSU: Optical Subscribing Unit
OTN: Optical Transport Network
OTT: Over The Top

P2MP: Point to Multi-Point
PAM: Pulse-Amplitude Modulation
PDCP: Packet Data Convergence Protocol
PHY: Physical
PON: Passive Optical Network
ppb: parts per billion

QoS: Quality of Service
Qxx: Question xx (xx=number)

RAN: Radio Access Network
RAT Radio access technology
RF: Radio Frequency

RLC: Radio Link Control
RoF: Radio over Fiber
RRH: Remote Radio Head
RRU: Remote Radio Unit

SDN: Software Defined Networking
SGxx: Study Group xx (xx=number)
SP: Signal Processing
sup: Supplement

TDD: Time Division Duplex
TDM: Time Division Multiplexing de-multiplexing
TDMA: TDM Access
TRX: Transmitting Receiving Things
TSN: Time-Sensitive Networking

UE: User Equipment

VLAN: Virtual Local Area Network

WDM: Wavelength Division Multiplexing de-multiplexing
WG: Working Group
WiFi: Wireless Fidelity

5. Conventions

None

6 Future Use Cases and Technology Drivers

6.1 Large capacity

According to [Cisco], the traffic in mobile communication networks is increasing at an annual rate of 61% and projected to grow 1000 times in the future. Therefore, it is required to summarize the issues as to whether the future requirements can be supported by the current network architecture for mobile communications.
Figure A.1 provides a VAN diagram outlining the requirements for future mobile communications. Compared with 4G, the future mobile communication requires larger capacity in the Extreme area, faster communication in areas such as Rural, Urban, Dense, etc. and expanded coverage in the isolated area [TTC].

Especially regarding the capacity increase, applications like AR (Augmented Reality) and real-time cloud access are assumed, with data rate requirements of 100 to 1000Mbps at any given time and around 10Gpbs at peak.

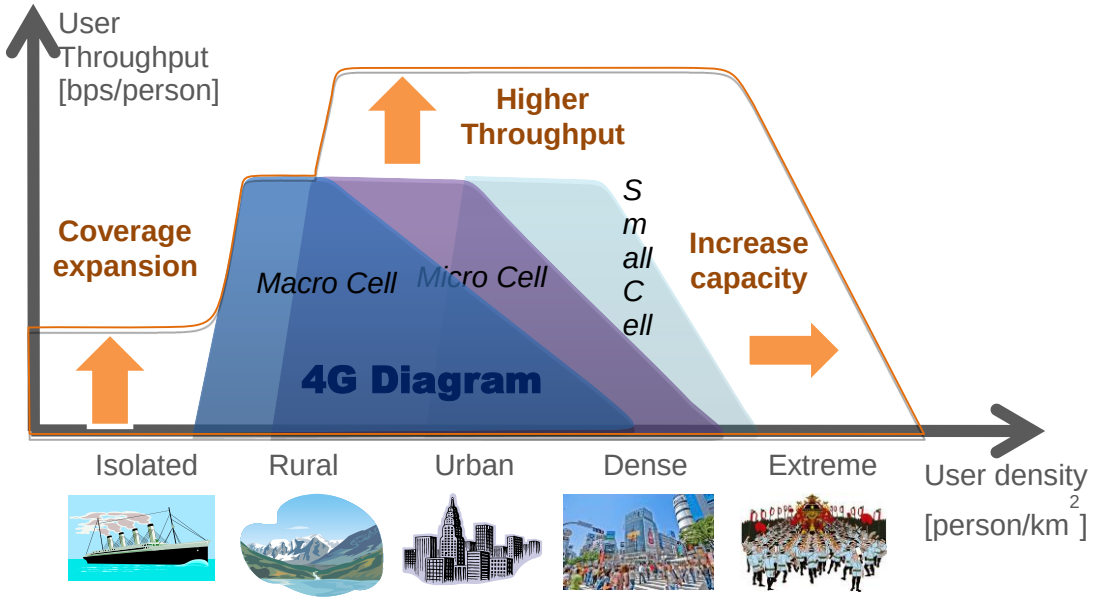


Fig. A.1. Requirements for future mobile communications

6.2 Low latency

In the future mobile network, it is expected that some new mobile services with very low latency requirements will appear, which could not be provided with 4G. Specifically, an E2E latency requirement of 1ms is being considered for such extreme applications as tactile communication, AR and auto-driving.

6.3 Power saving

The forecast traffic growth is very large (in bandwidth, in number of cells, in user devices). Continuing forward with the current state of the art power consumption levels would be unsupportable, both from carbon footprint and power availability perspectives. Therefore, the power consumption of all the elements must be controlled and reduced to avoid this problem.

6.4 Large-scale disaster/congestion/failure resilience

Disaster resilience can be considered from congestion and failure resilience perspectives.

For congestion resilience, the traffic during the Great East Japan earthquake in 2011 was 50 to 60 times higher than normal with regard to voice communication via cellar phones. Concentrated service requests from base stations that cover a wide area causes resource shortage and congestion. Telecommunication carriers then implemented 70 to 95% traffic control [MIC]. It was extremely difficult for users to establish a voice communication. According to a survey result, people made a call about 12 times on average until they succeed and about 14 times on average until they give up in disaster-stricken areas[MSR].

For failure resilience, with regard to unexpected communication process disruption due to damage of network functions, the earthquake and tsunami caused collapse, flooding and washout of building facility, split and damage of undergrad cables, duct lines, etc., damage of utility poles, damage of aerial cables and collapse and washout of mobile base stations, which resulted in severe damage [MIC].

Although no specific numerical target levels are shared as a future scenario in terms of disaster resilience, the government and users both demand further enhancement of

telecommunication networks based on the lessons learned from the Great East Japan earthquake described above.

6.5 Diversified types of terminal/traffic/operator

For the traffic for conventional mobile terminals used by people, as high-definition terminals with a large screen and video capture function become common, more and more various video contents are used as a medium and the OTT service is expanded, which increases video traffic. Furthermore, as M2M terminals get popular, the traffic of M2M terminals is expected to increase sharply.
In general, a connection topology like sensor network is assumed for M2M terminals, with possible use cases such as management, monitoring and remote control of production facilities, lifelines, building and housing, vending machines and heavy equipment. The device mobility is relatively low and both the occurrence frequency and data volume of each traffic tend to be small, but the number of terminal connections per unit area becomes very large. From 2020 and onward, along with the advancement toward IoT and IoE incorporating M2M, the terminals and applications to be accommodated will further diversify. It is also expected that there will be many new players in the mobile service industry as MVNO.

7. Technical Challenges

This chapter outlines some major technical issues for Mobile Front-Haul and Mobile Back-Haul networks.

7.1 Transport bandwidth

The rapid uptake of cellular data services is at the same time exciting and frightening. The rapid growth of traffic has very large revenue and business potentials, and operators are keen to take advantage of this. On the other hand, such exponential growth will soon out-strip the capability of 4G networks, even with the reinforcement that seems inevitable now. The usual incremental pace of improvements will not be enough. This is the rationale for the development of "5G wireless", which aims to provide 1000 times the bandwidth of 4G networks. Herein, we equate the "5G" class of wireless technology with that being considered in the IMT-2020 Focus Group.

This high-level goal of three orders of magnitude increase in capacity is often analyzed into three major enhancements. The first is an increase of bandwidth. While the usual unit of bandwidth in 4G is 20 MHz, that in 5G systems will be able to use 200 MHz; this enhancement should give directly a 10x increase in capability, if the network and terminals can terminate such a large bandwidth, and if such spectrum can be found on the airwaves [Comment] I couldn't understand this sentence well. Is this mean the following? "if such bandwidth can be allocated" [Proposed resolution] If yes, change the phrase to above one. . One would have to assume that if the demand is there, this will happen. [Comment] I couldn't understand this sentence well. Is it collect? [Proposed resolution] Please re-check the sentence.

The second enhancement is an increase of cell density. Beginning with the existing 4G macro-cell sites, we can imagine that 6~10 micro-cells would be placed around each. The increase in capacity is roughly an order of ten, but it is not so easy to calculate as the capacity of the micro-cells might not be as high as the macro-cells, and calculating "capacity" over a physical network is not a simple task. (For example, is it cell-edge rate, or aggregate average rate?)

The third enhancement is the exploitation of massive multi-input multi-output (MIMO) techniques. Included here are such things as antenna arrays at each site or sector, and coordinated multi-point (CoMP) transmission over a set of sites. To achieve the desired 10x increase in capacity, MIMO

orders of 64x64 are being routinely considered. This enhancement has a knock-on effect that to produce the MIMO gain efficiently (that is, to implement Coordinated Transmission and Reception), the centralization of the baseband processing is required. This has stimulated the interest in centralized radio access network (C-RAN) concepts, and hence wireless front haul.

Impact of these three enhancements on the optical transmission requirements should be considered. Increases of spectrum and of the number of sites will lead to direct linear increases of transport required. If a typical 4G sector is served by a 100 Mb/s backhaul link, a 200 MHz 5G sector would need 1 Gb/s. However, the introduction of MIMO and C-RAN leads to much more bandwidth.

First, front haul networks transport 30 bits per sample, while the actual information capacity in those samples is perhaps only 5 bits; this is a 6x increase from the back haul model. Second, the MIMO multiplicity directly scales the transport needed (e.g., 64x64 MIMO is a 64x increase in transport), because the data reduction occurs in a central location. Taken together, this amounts to a 400x bandwidth increase, so a single 200 MHz sector of 5G would need about 400 Gb/s of capacity. Combining that with 10x for network densification and thus resulting in 4 Tb/s of capacity for each macro-site sector.

Gap D.7.1-1: Large capacity transmission. See Clause 7.4.1 of the main body of this report.

7.2 Functional split

Clearly C-RAN bandwidth requirement needs to be addressed. Large bandwidth capacities are now seen in long-haul networks and rather expensive. The costs of this transport would outweigh the presumed benefits of the 5G wireless system. At this point, some more effective system engineering is needed to help balance the demands of the wireless with the capabilities of the optical transport system.

There are already some remediation methods that have been proposed. For example, the current interface de-facto standard (CPRI) has certain overheads that are not so efficient, especially when the interface is scaled to higher rates, and when payload compression is used.

This is being addressed in some of the newer interface standards being developed now. The number of bits per sample can be adjusted, or the samples can be compressed to some degree. These payload compression methods (which can be lossy) can achieve perhaps a 2 to 3x reduction in bandwidth demand, but they also come at the price of reduced signal fidelity. They also make the processes of signal shaping and automatic gain control more critical, as the bit depth of the system is being reduced.

Beyond these small factors, another approach would be to rethink the entire C-RAN architecture, such that some functionality is pushed out to the remote radio units. This can reduce the data volume by large amounts; however, it sort of reduces the utility of the system. It makes the RRU have a bigger size, power consumption, and cost. Moreover, it prevents a large part of CoMP gain. So, while the re-architecting of 5G might have to be used at some point, there is still an interest in keeping the remote simple and centralizing as much processing as we can.

7.3 Network Timing and synchronization

Stringent latency requirements (of the order of 1 ms) is one of the main target for some of the applications that need to be supported by 5G (e.g. automatic traffic control, remote surgery, tactile internet).

In this respect the following is stated in the [NGMN]:

"The 5G system should be able to provide 10 ms E2E latency in general and 1 ms E2E latency for the use cases which require extremely low latency."

Consequently, this imposes significant constraints on the transport network in order to meet such a requirement, particularly, in packet transport technologies, where queuing and processing delays over multiple hops could easily exceed the above limit.

Of this point it should be noted that initiatives already started within IEEE [TSN] and IETF [Detnet].

Network synchronization and distribution of accurate time and frequency synchronization references in the network is another key issue for a successful deployment of 5G.

Several aspects of network synchronization and distribution require careful analysis and below some of the relevant items are highlighted:

- Wider use of TDD (Time Division Duplex) as a radio technology. This is a main example where accurate phase alignment is required between radio frames in order to control interference between uplink and downlink signals delivered by adjacent base stations and/or UEs.
- Continued increased use of features implying time synchronization requirement such as carrier aggregation and broadcast.
- Dense radio base station deployments leading to potentially greater need for coordination.
- New Radio technologies (and related new numerology) may be defined (e.g. addressing some of the key 5G requirements) potentially implying requirements for more stringent than the 1 microsecond that is currently considered by ITU-T SG15 Q13.
- Sharing of resources will also likely require new paradigms in terms of administration and operations of the synchronization networks.
- Ongoing evolution of SDN and NFV concepts and related impacts on synchronization network architecture and operations.
- New applications, and users (IoT in particular), potentially leading to new synchronization demands.
- 5G requirements on transport in terms of reliability, latency, robustness (where the fronthaul segment is one of the most demanding use case from a synchronization), synchronization as a key enabler for transport.

In general, 5G will impose a number of requirements (such as higher capacity, latency, handling with a single platform a number of applications that require timing such as industrial automation, energy efficiency) that in one way or another could result in an accurate network synchronization requirements.

The network architecture is also particularly relevant from a synchronization perspective due to new concepts like NFV, SDN, Cloud, distributed applications that may also impact the way synchronization networks will be defined.

The main group within ITU-T dealing with synchronization is SG15 Q13.

This Question (in cooperation with other relevant questions in SG15) has been working in the last few years on solutions to deliver time and synchronization references over packet and Optical transport networks. Proper design is required in order for an optical transport networks (OTN) to meet the applicable synchronization requirements.

However some of the most stringent requirements applicable in fronthaul networks cannot be met by the existing standards.

However, more work is required to address 5G requirements and the stringent latency.

Gap D.7.3-1: Timing requirements. See Clause 7.4.1 of the main body of this report.

Gap D.7.3-2: Low latency. See Clause 7.4.1 of the main body of this report.

7.4 Power efficiency of fronthaul

Figures A.2, and A.4 show the configuration of Mobile Fronthaul.The major power saving issues in the Mobile Fronthaul are:
- The connection between the BBU and RRH always uses a fixed rate regardless of actual traffic volume.
- Deployment of small cells (increase of the number of devices) increases the total power consumption.

- Faster optical transceivers, electrical processing circuits, etc. between the BBU and RRH due to higher data rate over radio increases power consumption.

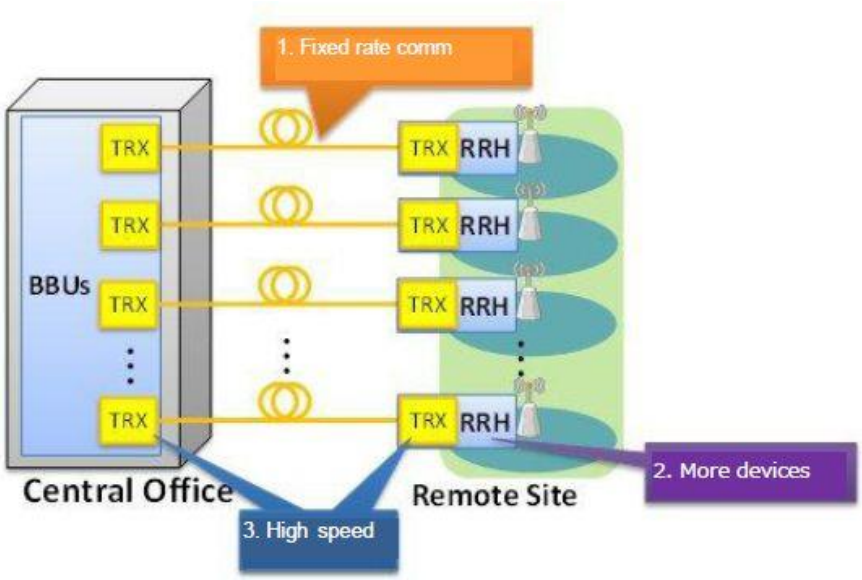


Fig. A.2. Mobile Fronthaul configuration and power saving issue

Power consumption due to fixed rate communication
As is the case in the Mobile Backhaul, mobile communication traffic fluctuations by time of day is also an issue in the Mobile Fronthaul. Especially, traffic fluctuations between cells are greater, which is likely to become more remarkable with small cell deployment. Therefore, the current standard and system design that use a fixed rate for communication causes wasted power consumption during the hours with light traffic.

Increase in total power consumption due to increased devices

For the future mobile network, small cell deployment is being considered to cope with further increase in traffic. This causes increased cells (i.e., increase in the number network devices that constitute a cell), which raises a concern about increase in total power consumption. Then, the impact of small cell deployment on the total power consumption for the Mobile Fronthaul has been estimated as follows.

Table A.1 shows the estimate assumption, which is typical of the network in Japan. It is assumed that there are 100,000 macro cell sites with 6 sectors for the current network. For the future network, it can be assumed that in addition to the current macro cells small cells are superimposed and there are 1 million 1-sector small cells. The equipment power consumption used for the estimate was determined in reference to the [Docomo] document. In addition, two types of small cell transmission rate (1Gbps and 10Gbps) were examined. At that time, the power consumption of 10Gbps was calculated as 1.5 times of that of 1Gbps.

Table A.1. Estimate conditions

	Current	Future
Cell config	MBH (Macro cell)	MBH (Macro cell) + MFH (Small cell)
# of cells	100,000	100,000 (Macro) + 1 million (Small)
# of sectors/cell	6 (Macro)	6 (Macro) 1 (Small)
Equipment power consumption	Macro: 4.5 KW (6 port BDE + 6 port BRE)	Macro: 4.5 KW Small: 1.2 KW (6 port BDE), 0.1 KW (1 port LRE)
Transmission rate	Macro: 1 Gbps	Macro: 1 Gbps Small: 1 Gbps or 10 Gbps

The power consumption for the current network (*Pcurrent*) and the power consumption for the future mobile network (*Pfuture*) are calculated as follows. Where, *Ncell* is the number of cells, *Nsector* is the number of sectors per cell, *Pequip* is power consumption of equipment, and *Nport* is the number of ports on the equipment.

$$\begin{cases} P_{current} = N_{cell}(macro) * N_{sector}(macro) * \left(\frac{P_{equip}(BDE)}{N_{port}(BDE)} + \frac{P_{equip}(BRE)}{N_{port}(BRE)} \right) \\ P_{future} = P_{current} + N_{cell}(small) * N_{sector}(small) * \left(\frac{P_{equip}(BDE)}{N_{port}(BDE)} + \frac{P_{equip}(LRE)}{N_{port}(LRE)} \right) (* 1.5) \end{cases}$$

As a result of the calculation, the total power consumption for the Mobile Fronthaul is up to 900MW for the future network, twice that of the current network that is 450MW. Note that the power-generating capacity at a nuclear power station is about 500MW per plant.

Increase in power consumption of equipment due to higher data rate

The major factors for the power consumption for the Mobile Fronthaul include the optical transceiver part, the electrical processing circuit part and RF amplifier. As a result of higher data rate over radio, these devices need to be faster, which leads to increased power consumption.

The power consumption of optical transceiver is as shown in Table.A.2. For the optical transceiver, enough power saving is implemented on the level up to 10Gbps and therefore the impact is small. However, on the level of 100Gbps, power consumption increases sharply and the impact cannot be ignored considering the power consumption with small cell deployment. It is required to consider the impact of higher data rate on power increase also for the electrical processing circuit and RF

amplifier. High frequency bands may be added in the future network, so the impact due to added frequency bands also needs to be examined.

Power consumption of optical transmission equipment with ultrahigh capacity

(1) Power consumption of optical transceiver

The current product level of optical transceiver is as shown in Table A.2. Each type of transceiver listed in the table supports transmission distance of 40km. The power consumption of the optical transceiver part in the transmission equipment is simply the required number of transceivers times their power For instance, to achieve 1Tbps will consume 150W, and with about 100,000 macro cells and a redundant configuration, the total power consumption would be:

150W * 100,000 * 2 = 30MW.

Table A.2. Optical transceiver types and power consumption

Transmission rate	Standard	Power consumption
Up to 1 Gbps	1000BASE-LH	≦1 W
Up to10 Gbps	10GBASE-ER	≦1.5 W
Up to100 Gbps	100GBASE-ER4	≦9 W
	Digital coherent	≦20 W (DSPonly)

(2) Power consumption of electrical processing circuit (interface process)

In addition, the power consumption of interface processing part of the existing switch equipment is about 30W per 10G-1 port. So the power consumption for the interface processing part to achieve 1Tbps is 3000W, which is 20 times greater than optical transceiver. And the consumption for the entire network is 600MW. Therefore, integration of electrical processing circuits (40G, 100G) is necessary to reduce the power consumption.

Gap D.7.4: Power saving by sleep or rate control. See Clause 7.4.1 of the main body of this report.

7.5 Large number of small cells

Figure A.2 shows the configuration of the Mobile Fronthaul. Due to high-speed data rate of mobile terminals (great capacity at a cell), the capacity of the line used for the Mobile Front-haul needs to be increased. For example, a transmission capacity of about 160Gbps (about 16 times) is required to support 10Gbps terminals in the current CPRI-based Mobile Front-haul.

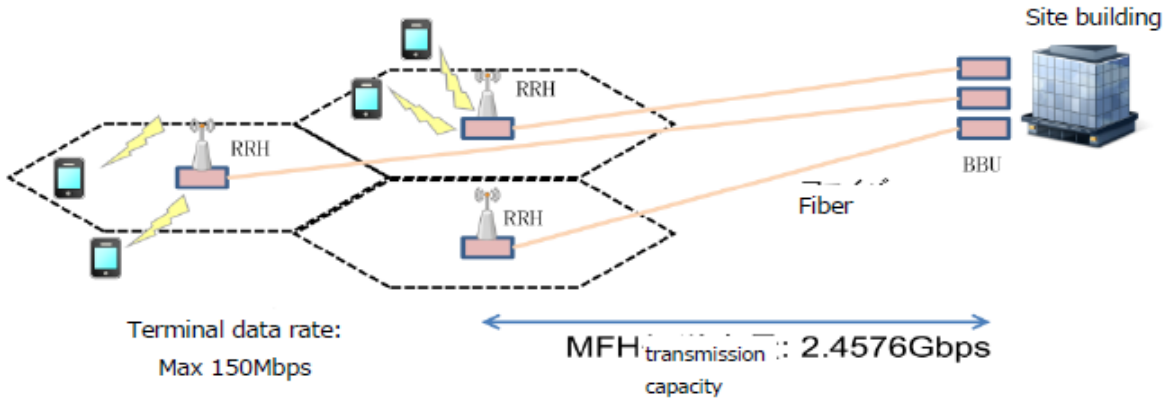


Fig. A.3. Configuration of Mobile Fronthaul

Furthermore, widespread deployment of small-size cells is expected to support high-speed and large-capacity mobile communications. In addition to macro cells with a radius of several kilometers, small cells with a radius of some dozens of hundreds of meters are being considered to be deployed together. For instance, assuming that a macro cell of 2km radius is replaced with small cells of 200m radius, the number of cells calculated based on the superficial area would increase 100 times. This brings up a concern about sharp increase of network cost due to increase in the number of links in the P2P configuration used for the current fronthaul.
Figure A.4 and Figure A.5 provide the number of links in the macro/small cell. If, for example, a macro cell (2km radius) is replaced with small cells (200m radius), the following are expected.

- The number of small cells increases 100 times.
- Required fibers and MFH optical transmission equipment also increase 100 times due to the increase in the number of small cells.

The cost increase due to large capacity of MFH optical transmission equipment needs to be taken into account.

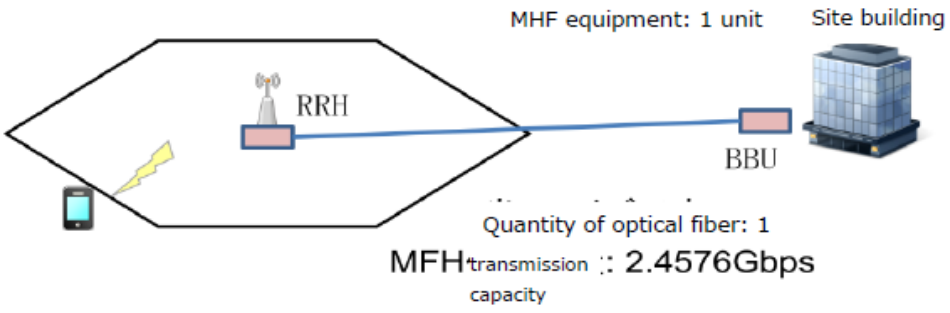


Fig. A.4. Number of links at macro cell

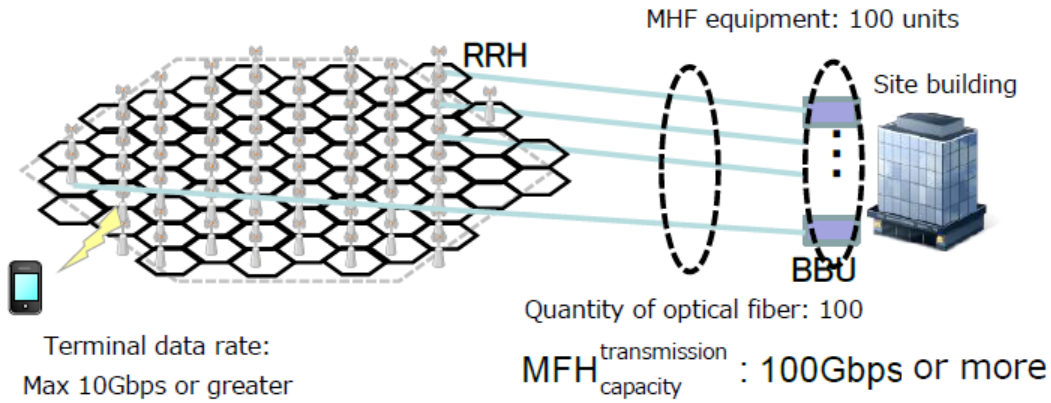


Fig. A.5. Number of links at small cell

Summarizing the issues for the Mobile Fronthaul based on the above discussion, the major issues are the followings: (1) Large-capacity transmission of 100Gbps or more and (2) Increase in the number of links.
(1) Regarding the large-capacity transmission of 100Gbps or more per cell, possible methods include reduction of transmission data amount and improved efficiency with transmission data compression. In the current CPRI transmission, however, radio signals in use cannot be identified at the optical layer, requiring all radio signals to be sent. The bandwidth in use also cannot be identified at the optical layer. So the calculation is made using the peak rate.

(2) Regarding the increase in the number of links, the number of fibers and equipment is expected to increase as long as the current P2P configuration is used, causing the cost to increase. Thus the system change to P2MP may need to be considered. A specific method for achieving this can be PON (TDM/WDM techniques, etc).

Gap D.7.5-1: PON as the virtual digital wireline service. See Clause 7.4.1 of the main body of this report.

Gap D.7.5-2: Large number of fibers for front haul. See Clause 7.4.1 of the main body of this report.

7.6 Reliability and resilience

In the future mobile network, increase in traffic to be accommodated and expansion of accommodated terminals including IoT are expected and the importance as social infrastructure will be greater than ever. Therefore, the network needs to be more robust than ever against congestion and fault in the event of disaster or fibre cut. The existing network can only cope with congested traffic during disaster by temporarily managing network resources and therefore does not ensure sufficient network resources necessary during an emergency. It is necessary to take fundamental actions for the future network such as allowing for prompt enhancement.

It is necessary to build a network with high reliability that can secure communication lines, flexibly and dynamically responding to Backhaul node change and topology change in such events as base station outage.

ITU-T SG15 developed a number of protection recommendations such as G.873.1, G.873.2, G.8031, G.8032.

Gap D.7.6: Reliability and resiliency. See Clause 7.4.1 of the main body of this report.

7.7 Diversified types of terminal/traffic/operator/FH&BH

The future mobile network is expected to further permeate society than the conventional mobile network. Not only the conventional terminals like feature phones and smartphones that assume usage by people, but also a number of terminals assumed to be embedded in devices are expected to emerge, creating a variety of equipment. As a result, the traffic pattern may also be different. The end point of communication will be machines instead of people, and the number of terminals for M2M communication is expected to increase exponentially. Furthermore, the information exchange in M2M is expected to have a traffic pattern that differs significantly from the server-client data exchange in the conventional IP network. In addition, a variety of operators are expected to operate mobile networks. Thus, new challenges for the network are generated by diversification of terminal requirements, traffic patterns and mobile network operators.

In order to efficiently accommodate numerous M2M terminals, it is considered to be necessary to implement policy control, addressing, etc. on a group basis for M2M terminals, which are different from normal mobile terminals.

Along with diversification of applications including M2M/IoT, the number of MVNOs is expected to increase to further improve the convenience of users. The network has to be flexible to release sufficient resources for necessary core network functions to MVNOs. In meeting a number of release requirements, there will be a various function requirements for each MVNO, which requires scalability to meet new requirements.

For mobile operators, transport might become more complex and more flexibility might be required to provide high quality services with reasonable cost, especially when networks deployment are becoming denser.

It is expected that in the next-generation IMT-2020/5G networks, many different types of base stations/devices are likely to be deployed, with different transportation requirements and targets. As shown in Fig. A.6 [ITU-R_F15], the transport in future IMT network would involve base station (BS) to device, device to device, and furthermore, BS to BS (or BS to dedicated relaying node) to transport the data traffic back to/from the core network.

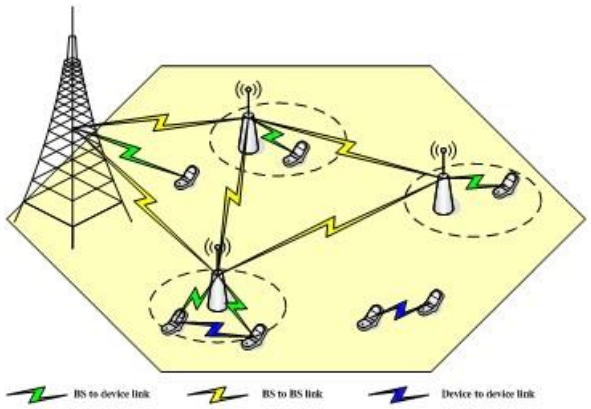


Fig. A.6. Illustration of different transport links in future IMT networks

Transport that is purely relying on optical fibre and microwave transmission may be inefficient and costly to provide the end-to-end transport service in the future dense deployment due to economic and/or propagation condition constraint. Wireless transport would be introduced for its inherent flexibility, low cost, and ease of deployment. It is therefore expected that the hybrid deployment, including optical fibre, microwave, wireless and other medias/technologies would be the case for BS to BS (or BS to dedicated relaying node) transport.

On the other hand, statistics show that in dense and heterogeneous networks, traffics of different BSs at different locations vary quite a lot, which is due to the non-uniform traffic distribution, and time-varying traffic that results in high peak-to-mean data traffic ratio at a given location. It in turn indicates that statistical multiplexing of radio resources become possible. By flexible use and assignment of the radio resources (including spatial-, frequency- and time-domain resources), the hybrid fibre/microwave/wireless deployment for BS to BS transport could meet the multiple requirements on achieving high capacity, while maintaining low cost and ease of deployment. Furthermore, the flexible use of radio resources among BS to device, device to device, and BS to BS transport might show more significant benefit to meet the different transportation requirements and targets with specific traffic distributions and propagation environments. Therefore, new transport solutions must be flexible enough to make sure that the scarce radio resources among the network could be multiplexed statistically to match the required service traffic distribution and the related propagation environments.

The flexibility requirement includes the capability of flexible topology and the capability of flexible resource assignment or sharing. The former refers to the capability of flexible use of spatial-domain resources, i.e., the deployed devices, and flexible configuration of the connection of the network nodes. The latter refers to the capability of flexible sharing and use of the time- and frequency-domain resources with the flexible configuration of the topology.

Besides, such flexibility needs also to improve other issues, such as reliability, co-existence with other solutions, fast deployment, support of multiple applications with different QoSs, network level energy efficiency, etc.

One example of flexible topology is shown in Fig. A.7 [ITU-R_F16]

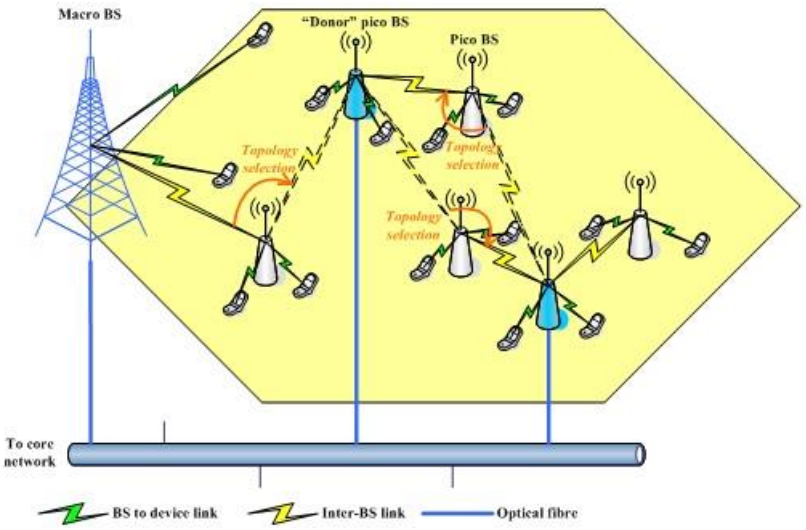


Fig. A.7. Illustration of flexible topology

Gap D.7.7-1: Diversified types of terminals. See Clause 7.4.1 of the main body of this report.

Gap D.7.7-2: Diversified types of traffic . See Clause 7.4.1 of the main body of this report.

Gap D.7.7-3: Diversified types of network operator. See Clause 7.4.1 of the main body of this report.

Gap D.7.7-4: Diversified types of RAN. See Clause 7.4.1 of the main body of this report.

7.8 Support of network slicing / management with FH&BH

One of the major architectural themes in 5G networking is the sliced network. The goal of this concept is to provide as much access to the underlying network capabilities to the upper layer applications. Front haul and back haul are part of the network, and hence are subject to this goal.

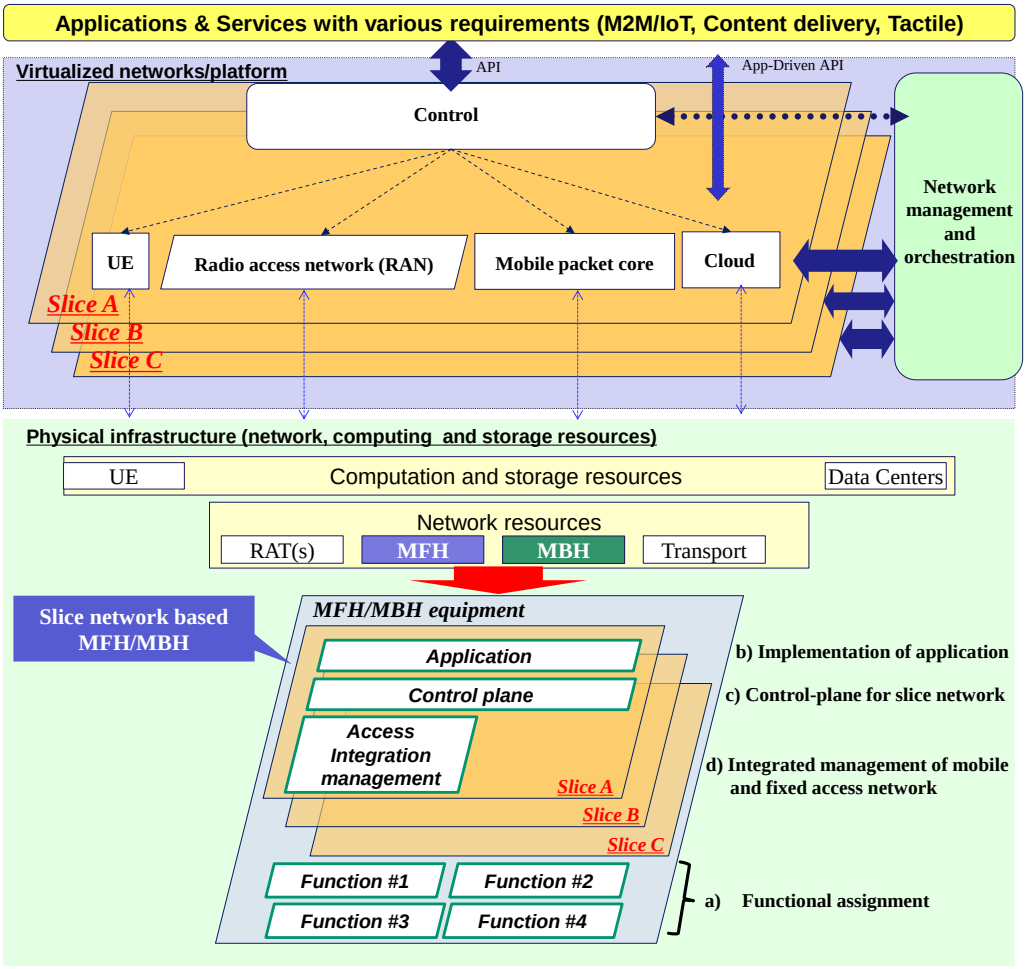


Fig. A.8. The sliced network and its relationship to MBH and MFH

a) Functional assignment at aggregation part of MBH and MFH
In the mobile network for 5G, introducing logical network, namely, slice network, which means separated network according to applications is assumed. In addition, transport system in core network and MBH/MFH should keep interconnectivity with the existing IP network technology where possible. The functional assignment at aggregation part of MBH and MFH is one of the key discussion points for smooth interconnection between core and MBH/MFH network in the sliced mobile network.

b) Implementation of application in MBH and MFH
For resiliency of mobile network, interworking between application and network is required. Implementation of application at aggregation part of MBH and MFH realizes flexible control according to use-cases and requirement of application by appropriate use of API, where possible. However, it must be noted that typical front-haul networks are transporting very low-level unresolved wireless signals, and so the control would be at an aggregate level.

c) Control-plane for slice network in MBH and MFH
For flexible control of mobile network, control-plane for slice network is required in MBH and MFH. The assignment of control functions can be configurable for requirements of each slice network.

d) Integrated management of mobile and fixed access network
Optical access technology is strong candidate for MBH and MFH. For construction of flexible and cost effective network, integrated management of mobile and fixed access network is required.
[Missing a gap here]

8. Architecture and Solutions for Mobile Fronthaul

8.1 Digital Radio (CPRI) over optical fiber

The simplest solution is to use simple point to point fibers to connect inexpensive data-com transceivers between the remote and central sites. The cheapest data rate at the current time is 10 Gb/s. The difficulty here is that to serve each 400 Gb/s sector would require 40 fiber pairs. This is obviously too fiber-hungry for a realistic network deployment. Currently, 25 Gb/s interfaces are starting to rise in popularity, and so this could reduce the fiber requirement here to 16. This is better, but still not so desirable. There is continuing work on even higher speed line coding, such as PAM-4 and 50 Gb/s transmission, which taken together could get us to 100Gb/s per wavelength. The relevant standards in this case would be those developed in the IEEE P802.3 group (LAN/MAN, a.k.a. Ethernet).

To reduce the fiber count further, WDM can be employed. For instance, 100 Gb/s interfaces that employ 4 wavelengths of 25 Gb/s each are commonplace now. This can get the system to 4 fiber pairs per sector, which starts to be reasonable. Using even more wavelengths can get us to a single fiber per sector, which is indeed an attractive design point from an equipment perspective. However, all of the WDM approaches suffer from the fact that while fiber is conserved the equipment still has a large number of opto-electronic components. Perhaps this can be miniaturized using silicon photonics, but for now it remains a more bulky packaging arrangement. From another perspective, it would be better to reserve WDM for the network multiplexing of multiple sites, and use some other method to develop the capacity for a single sector on a single wavelength channel. The relevant standards here would be the G.989 series (NG-PON2) and G.9802 (Multi-wavelength access systems), both developed by Q2/15; and the G.698.3 and G.metro recommendations, both developed in Q6/15.

If we set our goal at larger capacity per wavelength, then we must consider more spectrally efficient use of the channel. This leads us to the use of OFDM over the optical link. If we begin with optics with bandwidth typical for a 25 Gb/s, then the RF bandwidth is about 20 GHz. Employing OFDM could perhaps get us to 5 b/s/Hz, producing a 100 Gb/s link on each wavelength. This approach reduces the requirement for WDM but it doesn't entirely eliminate it. Also, if one considers what is happening in this system, an OFDM signal (5G wireless) is being digitized and then carried over another OFDM signal (optical). This double modulation is kind of inefficient from a processing perspective. Currently, there are no standards for this type of link. They may be developed in Q2/15, or potentially Q6/15; however, no formal plan has been established.

Gap D.8.1-1: Optimization of module or chip device design. See Clause 7.4.1 of the main body of this report.

Gap D.8.1-2: PON with WDM overlay. See Clause 7.4.1 of the main body of this report.

8.2 Analog Radio over optical fiber (P2P and PON)

Taking this thought to the next step, directly carrying the wireless signal over the optical link seems more efficient and natural. This is basically radio over fiber (RoF). In this case, the different MIMO channels would be multiplexed using frequency division multiplexing. Assuming we begin with a 20 GHz bandwidth optical link, frequency division multiplex can easily fit 64 channels of 200 MHz in this space using reasonable guard bands. So this system can achieve the goal of carrying an entire sector on a single channel. Unlike earlier RoF systems, this application is unique in that all these channels are being multiplexed in a single location. Hence, the necessary RF processing can be accomplished digitally, using DSP techniques. Making the job easier is that all the MIMO signals would belong to the same system, and therefore are identical in carrier frequency. The viability of this new kind of digital frequency multiplexing based RoF has recently been verified experimentally.

Of course, nothing comes without a drawback, and in this case, the analog optical transmission will induce some loss of signal fidelity. The major impairments here include quantitization noise in the DSP and DAC/ADC, noise and nonlinearity in the optical-electrical conversion, and optical channel impairments (e.g., dispersion). The back-to-back fidelity limited by the resolution of DAC's and ADC's can be engineered to 1~2% error vector magnitude (EVM), and over a reasonable link budget 3~4% EVM can be achievable. The main optical constraint is signal power, and so optical amplification is very effective in extending the reach, if required. This technology has been described in a Supplement to the G-series of ITU-T recommendations (G.sup.55). Work on a normative recommendation for some RoF systems has begun.

Using the RoF link as a building block, it would be possible to use WDM-PON technology to multiplex signals from a typical macro-site onto a single fiber. A hypothetical network could consist of a fiber (or fiber pair) running from the central processing point to the macro-site, where a wavelength multiplexer element would derive multiple drop signals. Some of these would serve the sectors running from the macro-site, and some would then be conducted to the surrounding micro-sites. An interesting question arises here: do we need so-called colorless technology? WDM-PON has always supported the idea that the end stations would be colorless and automatically tune to the right color determined by their network connection. In this 5G wireless application, the number of deployed units is far smaller than in typical access, and colorless operation is perhaps not necessary.

Gap D.8.2: Analog radio over optical fiber transmission. See Clause 7.4.1 of the main body of this report.

8.3 Digital Radio (CPRI) over optical transport (OTN)

CPRI (or similar) interfaces can be transported over the Optical Transport Network (OTN) system. ITU-T SG15 recently agreed a new supplement G. Suppl 56 - Transport of CPRI signals in optical transport networks (OTN).This supplement describes alternative for mapping and multiplexing CPRI client signals into OTN. These mappings include direct mappings for native CPRI client signals and mappings that apply transcoding in order to gain bandwidth efficiency.
However, there remain issues regarding symmetry requirements and network timing performance, as it is very demanding to deliver the very stringent (2 ppb) frequency accuracy over normal OTN systems.

Gap D.8.3: CPRI over OTN. See Clause 7.4.1 of the main body of this report.

8.4 New digital format replacing CPRI

The CPRI defines a very simple data link that was originally intended to operate over short (~100m) fiber links running at ~1.25Gb/s rates. Its design was not optimized for use over a larger network. Some examples of the inefficiencies in CPRI include:

- It reused elements from 802.3 PHY clauses to define its optical and line coding methods. For instance, 8b10b coding is used for rates 10 Gb/s and below, which is a 20% overhead.
- The actual RF data (I and Q samples), CPRI also carries OAM information, at a fixed ratio of 1:15. While this was suitable for its lowest data rates, it becomes increasingly oversized for higher rates (e.g, 10 Gb/s CPRI has ~600 Mb/s of OAM throughput!?)
- The CPRI link is intended to be used for "line timing" of the RRU, and this, coupled with the very strict radio regulations on channel assignment, makes the CPRI timing requirements very strict.

Thus, some considerable improvements could be had if CPRI was revised to update its design to reflect its new intended application.

Gap D.8.4: Improved CPRI. See Clause 7.4.1 of the main body of this report.

8.5 Radio over Packet

Current radio over X systems all use a circuit-based transport paradigm. It would be useful if the radio information could be carried over a packet-based system, such as Ethernet, MPLS or IP. If this were possible, then all of the different packet transport solutions could be used, giving the network operators many more options.

In LTE, packet network devices are already widely used in the backhaul. Although in the fronthaul, CPRI has more stringent delay, jitter and synchronization requirements than that in the backhaul, but some new services in 5G may also have a very low E2E delay requirement, such as 1ms. In 5G, fronthaul and backhaul may have similar requirements, and it would be useful to have an integrated fronthaul / backhaul, carrying fronthaul traffic, backhaul traffic and even other types of traffic such as IoT and WiFi traffic in a same network. An integrated fronthaul and backhaul will simply network management and reduce OPEX.
The major issues encountered when sending radio data over packet networks include:

- Fragmentation and encapsulation of the data stream
- Circuit emulation in the face of typical packet impairments
- Maintaining all the timing requirements simultaneously

There is some work to address these issues. The IEEE 802.1CM will develop a profile for Fronthaul over Ethernet bridges. The IEEE TSN project proposes some solutions for precision timing and reduced delay and jitter. The IEEE 1904.3 project is devising a Radio over Ethernet encapsulation standard; however, that is the limit of its scope. The Next Generation Fronthaul Interface (NGFI) group is just beginning work in this area.

Gap D.8.5: Radio over packet. See Clause 7.4.1 of the main body of this report.

8.6 Function Splitting in MFH

Re-allocation of the functions between the base station and the remote antenna site can reduce the capacity required in MFH. Several function split points are under consideration as shown in Fig. A.10. When the function split point is defined in a higher layer (at a more left point in the figure), the required capacity becomes smaller, but it becomes difficult to realize Coordinated Multi-Point (CoMP) transmission/reception.

The following options are possible split points:

(a) CPRI (conventional)
(b) Split PHY
(c) MAC-PHY
(d) Split MAC
(e) RLC-MAC
(f) PDCP-RLC
(g) Service

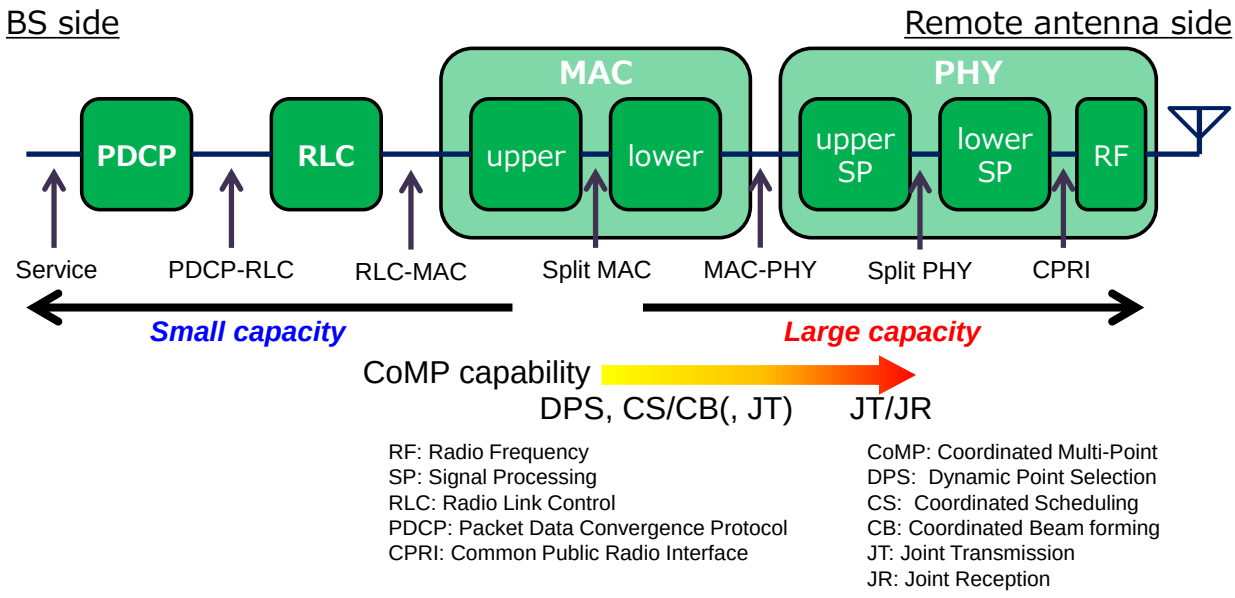


Fig. A.9. Options of function split and required capacity

To realize the future MFH with a new function split discussed above, we need a new signal format, i.e. a frame. That can not only reduces the capacity in MFH compared with CPRI but also allows various wire-line networks to be used as the base for the MFH. For example, Ethernet frame is one of the candidates.

Gap D.8.6: Developing function splitting of front haul network. See Clause 7.4.1 of the main body of this report.

8.7 Reuse of existing access networks

At present, broadband access with capabilities over 1 Gb/s are widely deployed in several countries. The major relevant systems are G-PON, GEPON, XG-PON, and 10GEPON. These operate using a TDM/TDMA scheme to share a single optical wavelength channel, and the systems provide generic packet transport.

Of course, it would be beneficial to reuse as much of this infrastructure as possible. Using the actual TDMA PON transmission system would require radio-over-packet style of interworking, mentioned above. Even if that is not done, reusing the same fiber infrastructure would be good. Overlaying wavelengths would be an example of this style of access network reuse. All of these aspects have already been described in ITU recommendations, and so no standardization gap seems to be present.

8.7.1 8. x Digital Radio (CPRI) over metro WDM (G.metro)

CPRI (or similar) interfaces can be transported over metro WDM system. ITU-T SG15 Q6 is currently developing a recommendation G.metro, where a WDM technique transport systems defined for metro applications. G.metro is targeted for applications where multiple services are delivered in a carrier network. Initially, tuneable laser is used and later coherent technology may be supported.

Metro WDM (G.metro) network offers great bandwidth and capacity in addition to increasing single port bandwidth and the wavelength counts that are multiplexed. Furthermore, the interface bitrate can be 2.5G, 10G and 25G with reach up to 80km. For the bandwidth capacity, G.metro offers optical wavelength grid of 100GHz and 50GHz spacing, resulting in 40 channels and 80 channels respectively. In the future a narrower grid and greater channel counts could be achieved.

CPRI could be transported transparently in G.metro network without electronics encapsulation and framing to meet strangent timing characteristics required by the mobile fronthaul network. G.metro offers smallest latency and could provide optical layer protection mechanism service resiliency. Furthermore, port-agnostic function could simplify the configuration and OAM of access layer in metro networks, and reduces the Capex and Opex.

G.metro supports horseshoe, chain and star topologies with WDM port-agnostic and independent upgrade of every port/wavelength.

Gap D.8.7: Extension of G.metro for the transport of CPRI in MFH/MBH networks. See Clause 7.4.1 of the main body of this report.

9 Solutions for Mobile Backhaul

A backhaul network is fundamentally a simple packet-based data aggregation network. In the conventional design of backhaul networks, the transport layer can be any number of packet transport technologies, including Ethernet, WDM, OTN, PON, and wireless. All of the established methods for management and control of such networks would continue to apply here, and there are few gaps foreseen for MBH. The following sections highlight a few open areas where issues still exist.

9.1 Network timing and synchronization in MBH

The MBH network must often provide a very accurate frequency and time reference signal. Fortunately, there are several methods that have been developed to support this, most salient being the IEEE 1588 Precision Time Protocol, and synchronous Ethernet. These methods can provide time and frequency to a usable level of accuracy, and are in use today.

What is missing is the allocation of the overall timing budget over the network as a whole. The MBH network is only one link in a much larger network. The IMT-2020 network has some overall limits for timing performance, but it is not clear how much of these limits can be allocated to the backhaul portion of the network.

9.2 Energy saving methods in MBH

As outlined in section 7.5, the power consumption of wireless networks will grow alarmingly if measures are not taken to reduce them. In the case of backhaul transport, the most direct way to reduce power consumption is with component technology improvements. Beyond that, the system can take advantage of the fluctuations in backhaul utilization to save energy. This can be done in several ways, as illustrated below.

9.3 Joint radio transport optimization

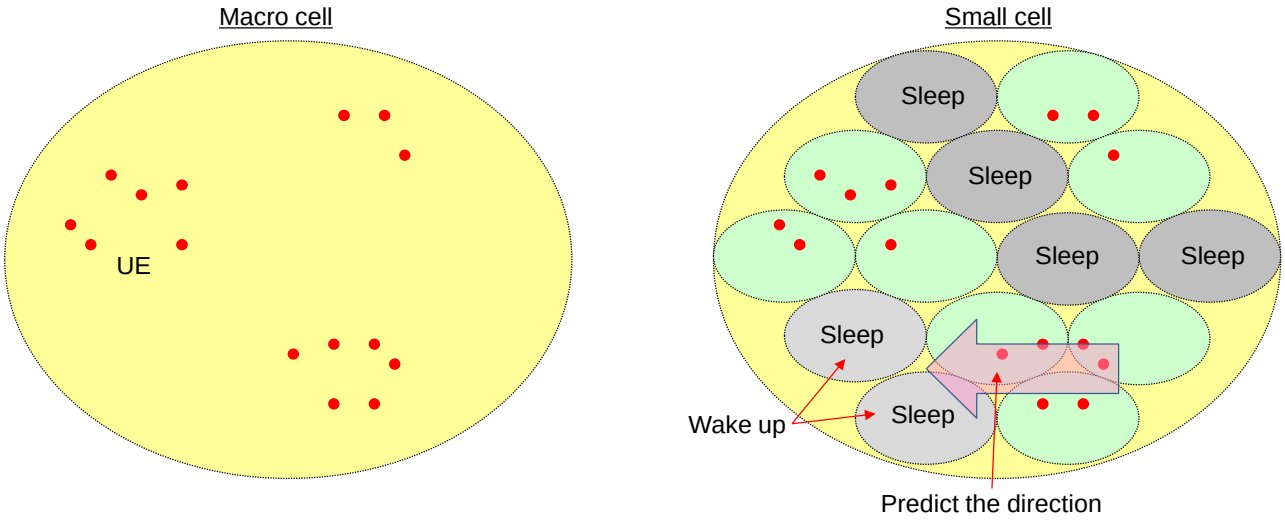


Fig. A.10. Sleep control of small cells associated with mobile

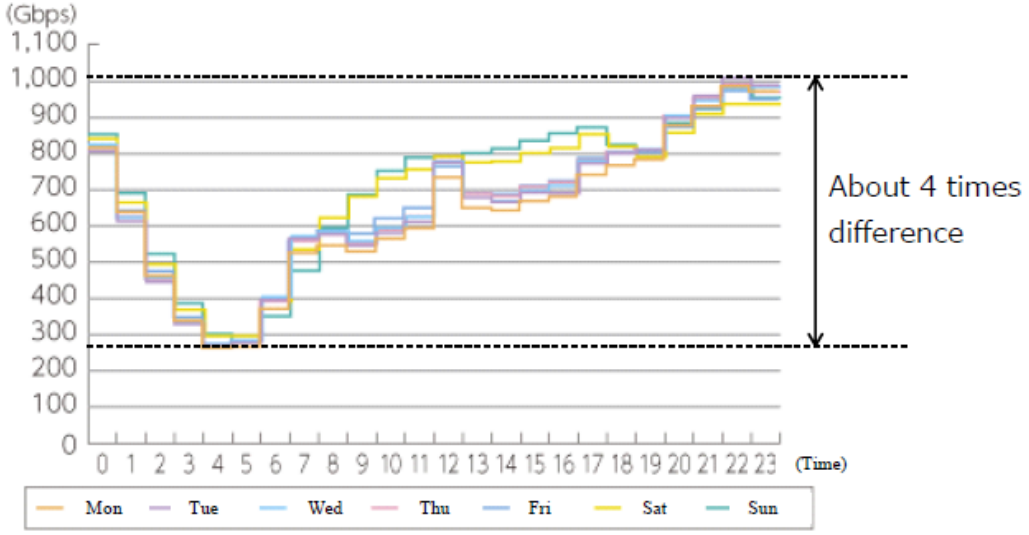
There is a case in which the terminal does not exist in the cell, since the area of the cell is reduced. Therefore, always driving the cell is a waste of power. Thus, the guess function of the mobile and sleep control are required.

Gap D.9.2.1: Coordination of power saving across MFH/MBH/Radio System. See Clause 7.4.1 of the main body of this report.

9.4 Adjusting transport to follow traffic fluctuations

Figure A.11 shows fluctuations of mobile communication traffic by time of day. Actual traffic greatly varies depending on the hour of day, with about 4 times difference between the max and minimum according to the current statistics. Therefore, if operating at the max transmission rate all the time, power ends up being wastefully consumed. By varying the number of operating transport channels according to traffic capacity, power consumption can be reduced. In the case shown

below, perhaps about 40% of the transport power consumption can be saved on average.



Source: Ministry of Internal Affairs and Communications, White paper on telecommunications for 2014

Fig. A.11. Fluctuations of mobile communication traffic by time of day

Gap D.9.2.2: Power saving by resource optimization (See clause 7.4.1 of the main body of this report.)

Appendix V

Emerging Network Technologies

Editor's Note: Appendix V was produced during the FG-IMT 2020 focus group in order to investigate gaps in standardization related to IMT-2020. While the request from SG-13 was to deliver a report outlining standardization gaps, the consensus of the focus group was that the working documents produced and used during the focus group work contained useful information for future work and should be captured. Note, however, the focus group concentrated on producing accurate descriptions of the standardization gaps in the main body of this document; some minor errors may exist in the appendices. They are, however, the output of the focus group but are provided for information only.

Editor's Note: This appendix uses clause references in a form usually associated for normative text. This is maintained for this report to align with references made in the main body of this report.

Draft deliverable of ICN for IMT-2020 Networks Working Group

1 Scope

This report explores the technology area Information Centric Networking (ICN) in the context of its use in the IMT-2020 network and its potential for assuring that the IMT-2020 network meets its visionary goals. The purpose of this paper is to identify the gaps related to ICN as an emerging technology to guide the future studies by ITU-T Study Group 13.

2 References

<to be updated with ICN applicable references (if any) inserted>

The following ITU-T Recommendations and other references contain provisions which, through reference in this text, constitute provisions of this Recommendation. At the time of publication, the editions indicated were valid. All Recommendations and other references are subject to revision; users of this Recommendation are therefore encouraged to investigate the possibility of applying the most recent edition of the Recommendations and other references listed below. A list of the currently valid ITU-T Recommendations is regularly published.

The reference to a document within this Recommendation does not give it, as a stand-alone document, the status of a Recommendation.

[ITU-T Y.3011] Recommendation ITU-T Y.3011 (2012), *Framework of network virtualization for future networks*.

[ITU-T Y.3012] Recommendation ITU-T Y.3012 (2014), *Requirements of network virtualization for future networks*.

[ITU-T Y.3033] *Framework of data aware networking for future networks.*

[Y.supFNDAN] *Revised Draft of Y.supFNDAN – supplement to Y.3033 on scenarios and use cases of data aware networking* (July 13-25, 2015, Geneva).

The following ITU-R Draft Recommendations describes the to-date architecture of the trasnport network of an IMT, including the radio network, the functions within a basestation, and between different base stations and the mobile infrastructure.

[ITU-R M.IMT.ARCH] Working Party 5D (25 June 2015), *Architecture and topology of IMT networks*

3 Definitions

Editor's note: this clause is currently empty

4 Abbreviations and acronyms

This Recommendation uses the following abbreviations and acronyms:

CCN Content Centric Networking

NDN Named Data Networking

FIB Forwarding Information Base

PIT Pending Interest Table

CS Content Store

ICN Information Centric Networking

5 Conventions

None

6 Motivation for Information Centric Networking

6.1 Overview of ICN

Information Centric Networking (ICN) is a different approach to addressing and framing data than today's Internet Protocol (IP) semantics. In IP, one uses a source and destination address to identify the two endpoints of a packet. The destination is almost always a unicast address and in a small number of cases an anycast address; the use of IP multicast is very limited. Inside the network, the payload of an IP packet is usually an arbitrarily framed byte stream (TCP) or datagrams (UDP). TCP/IP assigns an ephemeral name to each packet: source IP, source port, destination IP, destination port, byte offset, byte length. These names are not reusable, nor cacheable beyond use for retransmission of lost packets. ICN's approach is to assign a re-usable name to each packet or small group of packets. This allows object re-use and peer-to-peer messaging via name without needing to resolve endpoint identifiers beforehand. ICN also bundles object authenticity with the network packets, such as via Merkel signing of a group of packets in a manifest, so provenance stays with the objects even if cached.

There are several ICN architectures in active use today. The most widely known is Content Centric Networking (CCNx) and its offshoot Named Data Networking (NDN). NDN forked from CCNx around 2012. While there are several important protocol differences between NDN and CCNx, they are close enough in function that we will only describe CCNx.

Because ICN does not require resolving endpoint identifiers before using a name, it opens new possibilities in machine-to-machine and IoT applications. Today, IP-based applications must use specialized rendezvous mechanisms, such as link broadcast, multicast, dynamic DNS, multicast DNS, or SIP. This is because they must resolve an IP address for a desired name. ICN technologies remove the IP abstraction so the network can operate at the name level. This can make the network more responsive to application demands with less infrastructure.

Within a IMT-20205G RAN, ICN could serve as the object transport for intra-RAN data. For example, the state of a Slice could be stored and transported as ICN objects, so as services move between enodeB sites its state follows in the named ICN objects. ADD MORE POSSIBILITIES.

In the ICN technology CCNx, the name combines both a locator and identifier in to one routable hierarchical structure. One could think of it as routing on URIs, where each name segment can be arbitrary binary data not restricted to the URI syntax. At one end of the spectrum are pre-generated content names, such as for a movie. A movie service could name content with a prefix like /movie_service/superman/h264/768kbps/32kbps/English to indicate a codec and encoding rate. Names can identify things beyond static content. A simple example would be a dynamic web service, such as /book_store/home/<encrypted_account_identifier>, where the <encrypted_account_identifer> is a blob that the book store server can understand and use to generate a custom home page. Names could also indicate a type of calculation, for example /calc/4/2/times could return a content object with the value "8". In all these examples, we used ASCII names, but in practice name segments can be binary values not necessarily human-readable.

<Introduction to ICN – motivation for a change to ICN-based communication model; relevance to IMT-2020>

Figure 1 Draft architecture of 5G mobile network

6.2 Elements of ICN

An Information Centric Network is usually made up of content producers, content publishers, content replicas, and content consumers. A producer generates a piece of content, such as a document, photo, movie, or web page. It may have its own digital rights management (DRM) attached by the producer. A publisher packages a piece of content for use in the network. This may include pre-encoding the content to certain formats and names and signing them with a network identity. A replica distributes content from a publisher. A consumer fetches content via network names from replicas. The download process at a consumer understands the inherent security offered by the ICN, which usually allows authenticating every packet via direct signature or implicit hash chain from the publisher. This is different than today's security model, where authenticity derives from a secure connection to a replica. In the simplest configuration, one entity is a producer, a publisher, and a replica for its content.

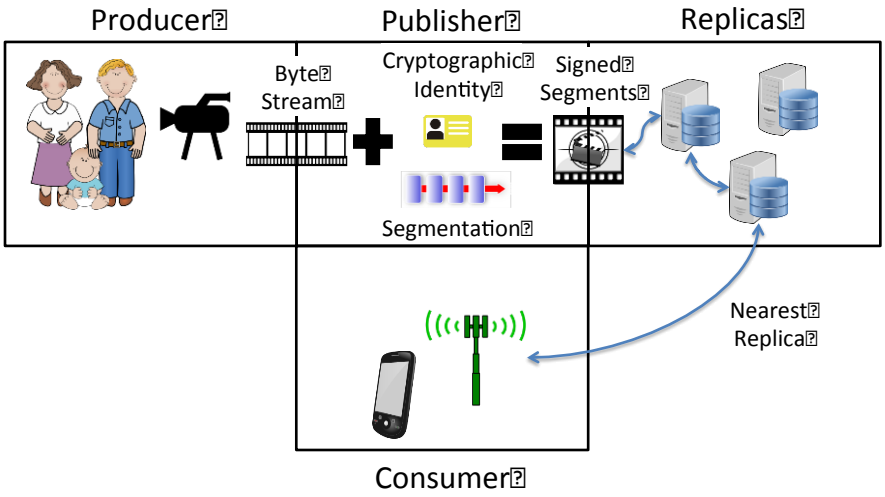


Figure 5. Typical ICN architecture

Figure 5 illustrates the typical ICN architecture, which we will make concrete by describing how an actual instance of CCNx would handle these activities. In this example, a family videos an activity at home, such as their baby. The video camera produces a structures MP4 byte stream. The publisher function – which may reside on the camera, home gateway, or other device – segments the byte stream to CCNx Content Objects. For a live stream like this, the publisher would segment it to a certain number of video frames in some number of network packets (Content Objects). A CCNx Manifest tree incorporates those Content Objects by hash in to a single signed manifest representing the whole video segment. The video segment is stored on a first replica, such as a home gateway. The publisher updates the movie catalogue to include the new segment, then repeats for the next segment. A consumer queries its nearest replica for the movie catalogue and segments. If the nearest replica does not have it, the request is forwarded towards the publisher until satisfied. The content travels to the consumer, optionally cached at intermediate replicas.

As illustrated in Figure 5, a consumer may fetch the CCNx Content Objects from any replica and still be assured that it is the correct data. This is because the Manifest tree is signed by the publisher and then securely hash-linked to each data Content Object. The consumer and replica may also opportunistically encrypt their session for privacy. The consumer may choose to only trust replicas, for example, that are enumerated by the original content publisher or are provided by the a trusted party, such as the user's carrier or cloud service.

In a second example, a cell phone producing a video could be producer, publisher, and replica all in one. Because one would not want a large number of consumers on the Internet fetching data directly from one's cell phone, it could be configured to only allow the user's home media server to fetch content and then act as the authoritative replica for the Internet. The user could choose to use a carrier service (i.e. cloud-based media server) to act as the authoritative replica.

6.2.1 Requirements relating to ICN in IMT-2020

The 5G network architecture should yield a framework that allows the reshaping of the economic foundation upon which the mobile network operates such that the new network can both efficiently serve the emerging use cases and markets (eg IoT) and support current usage models (eg those that

are video related) at substantially lower cost points. Our expectation is that properly crafted, the impact of a solution that successfully addresses these challenges would extend well beyond the bounds of the mobile network. In the following paragraphs, we identify characteristics of the new solution space that address the perceived deficiencies in current network design:

(1) Mobility – current networks base packet forwarding decisions on location-based labels that identify the point-of-attachment of the destination host on the network. Whenever the host point-of-attachment changes (ie a mobility event), a means for updating the association of the host to its network-addressable label is required, one that also preserves the socket-based connections that mobile applications are bound to. Mobile networks employ tunnel-based solutions <ref GTP> for this purpose. Tunneling ensures reachability of mobile hosts via a layer of indirection that allows the mobile host point-of-attachment label to change while preserving the socket structure. One can view the necessity of anchoring gateways required to implement and dynamically manage tunnel-connections as a direct consequence of network design that builds on a foundational layer that was not architected to support mobility. The current design burdens mobile traffic with an incremental cost that is traceable to this architectural consequence, one that requires mobility be treated as an explicit feature rather than as an emergent property of the basic design. The 5GIMT-2020 network should remedy this deficiency and treat mobility as a first class citizen in its design and eliminate the need for tunnelling and the associated complexity and cost that accompanies it.

(2) Privacy and Security – Privacy has always been a major concern for mobile networks. The RAN has employed link-encryption by default since the second generation standards and LTE commonly employs IP SEC in backhaul network between the eNB and PDN gateway for protection along those connections. The importance of guaranteeing privacy on all Internet services goes beyond the specific radio mobile data network or the all IP core; it includes the communication service to all of its components: over-the-air transmission, end-to-end IP, HTTP and web service that might employ user trackers or store private content in clear text at the server end. The current solution guarantees authentication, integrity and confidentiality by delegation to the transport layer on an end-to-end basis using TLS which may serve a short term objective but at a cost of significantly impairing long-term flexibility, reliability and manageability. The use of TLS implies a significant cost that is quantifiable[27] although it is the default solution in the HTTP/2.0 standard despite the criticisms about its weaknesses as a technical solution to the privacy problem.
The issues regarding user privacy and data authenticity, integrity and confidentiality are crucial to get defined accurately and addressed appropriately in the IMT-2020 basic design. It is important to recognize from the outset that different applications, usage models, market models, and so on will have diverse risk exposures that beget different security requirements. It is unlikely we can successfully anticipate all potential combinations and it is similarly unlikely that solutions that trend towards worst case coverage will be economically acceptable. This suggests an architectural design that enables flexible application of security elements appropriate to the application and usage requirement. The decomposition of security in component functions (eg data authenticity, integrity, confidentiality) such that each can be individually addressed is a desireable capability that allows the flexible creation of purpose-directed security and privacy solutions that can address differing risk profiles successfully within a cost-effective ecomomic model.

(3) Transport Efficiency – Reliable and stable transport over current networks is managed by a distinct protocol layer (eg TCP, SCTP, etc) that operates between communicating end-points. The end-to-end action of these protocols treat the composition of links that form the

[27] D. Naylor et al, “The Cost of the “S” in HTTPS”, *Proceedings of ACM CoNEXT, 2014.*

communication path in aggregate that lead to the use of abstractions for congestion detection and remedial action that lead to poor behaviour when applied to composite paths that include wireless links. The anomalously erratic behavior exhibited by TCP in mobile networks is traceable to the (mis)interpretation of variations in link bandwidth and latency as a sign of network congestion resulting in inappropriate backoff responses, unnecessary data retransmissions, increased download times and more generally inefficient usage of wireless access and network resources. An architectural approach that enables hop-by-hop congestion detection and management for accurate interpretation of link behaviour and initiation of appropriate responses is desireable. This is particularly important in the case of wireless links that employ (multiple access) link layer protocol may are highly adaptive and designed to deliver high spectral efficiency. The objective is to have intrinsic congestion and flow management mechanisms that account for the unique characteristics of the wireless link.

Current tunnel-based mobility transport also frustrates the use of in-network storage and caching of content within the mobility network. Caching should be viewed not simply as a means for improving delivery efficiency of popular content but also in its roll in facilitating retransmission and reducing latency when reliable delivery is required. The 5GIMT-2020 mobile network design should allow the flexible use of in-network storage as an architectural component.

(4) Latency Sensitivity: 5GIMT-2020 places stringent requirement such as 1-10ms for certain class of applications; also in general any applications benefits from reduced response time through better throughput and faster service logic execution. Though delays such as 1-10ms may be very difficult to achieve in software driven implementations, in normal situations distance of the consumer from the service contributes significantly to the end-to-end application delay. Though there are also tradeoffs to manage the distributed service instances, benefits are also realized as overall load of the service is amortized among the distributed instances. ICN features such as naming, name-based service discovery, name-based routing allows application to discover the closest service points with which it can transact. In certain situations, like IoT, these services can be placed at extreme edges of the network such as over public infrastructure to address the need of mission critical applications.

(5)

(6) Bandwidth Efficiency: The estimated increase in wireless capacity is expected to be 1000X for 5GIMT-2020. ICN can help this situation in multiple ways: 1) eliminating redundant data transmission considering multicasting and caching is integral feature of ICN; 2) cheaper computing and storage will enable significant intelligence end point and infrastructure elements well suited for ICN to exploit, here ICN enabled application and service interaction can be localized operating over both wired, unlicensed and licensed spectrum, thus offloading the backhaul from any control or data plane overhead due to these interactions ; 3) also tighter integration of ICN with MAC layer to enable multicast and broadcast of data objects can improve bandwidth efficiency of the wireless resources, particularly helpful for very popular contents and flash crowd situations.

6.2.2 Design Goals (expected functionality, usefulness of the components)

- object security
- latency minimization
- hop-by-hop dynamic congestion control

- multipath transport

- optimized transport reliability

- mobility

- …

6.2.3 Recent Research/Experimental Results

<summary of recent research results that are relevant to IMT-2020>

- Mobile Backhaul optimization

- Mobile Congestion Management

- Content Distribution

- …

7 Use Cases

- <principle applications of CCN in IMT-2020 eg IoT, popular content, multicast, network resource optimization, latency reduction, resilient networks/disaster recovery etc>

7.1 ICN- IoT

The "Things" referred to in an IoT framework belong to multiple scenarios and context making it difficult to have a unified set of requirements that spans all the scenarios. However, an attempt has been made in [28] to capture IoT requirements, though the priority of these requirements may be different depending on a given scenario and its specific context. We summarize these general IoT requirements and discuss why ICN meets these requirements. This discussion is also relevant considering many recent proposals in the form of AllJoyn, IBM ADEPT, Google-Thread, and more mature ZigBee/Zwave standards which tries to achieve one or more these requirements for specific scenario like home network, hence doesn't consider scenarios such as large scale mobility , or V2V scenario where the architecture has to meet the disruption tolerant requirements of Ad Hoc communication.

From an ICN-IoT implementation perspective several possibilities exist: 1) considering IoT currently is highly fragmented with multiple protocol suites, inter-operability is a major concern. ICN can be the unifying L3 protocol for the IoT operating because of the flexibility it offers in operating over heterogeneous $L2^4$ interface making the case of an end-to-end ICN implementation, the core ICN implementation itself may be overlaid over IP; 2) the other feasibility is using a proprietary/standardized protocols such a Zigbee instrumented for simplicity and energy efficiency in constrained segments, and applying ICN gateways to aid with information processing, publication and consumption.

7.1.1. ICN-IoT Design Goals and Gap Analysis.

---

[28] ICN based Architecture for IoT- Requirements and Challenges., " https://tools.ietf.org/html/draft-zhang-iot-icn-challenges-02 ", *IETF/ICNRG* 2015.

Here we discuss ICN-IoT requirements and gap analysis for each considering current research status. In general it has to be understood that, ICN-IoT is an emerging area of research within the ICN research space without any benchmark comparisons, hence the GAP analysis is presented as open research challenges

Naming and Name Resolution: The first step towards realizing a unified IoT platform is the ability to assign names that are unique within the scope and lifetime of each device, data items generated by these devices, or a group of devices towards a common objective. In ICN-IoT, we assign a unique name to an IoT object, an IoT service, or even a context. These names are persistent throughout their scopes. These names are resolved to locations of these named entities by the network as applications request access to them; this is termed as name resolution.

Research Gap : Naming is the fundamental design issue in ICN as it has to meet application requirements, with direct implication on the design of the ICN network layer and name resolution system. In the context of ICN, heterogenous radios and MTU restrictions offer another challenge of naming the constrained devices such as sensor and actuators. Further considering hierarchical nature of IoT deployment name translation may be required between local and global names for end-to-end IoT solution enablement. This topic is an active area of study [29] and challenges have to be addressed considering specific scenarios. Unlike static content in the general Internet, data evolves contiuosly in IoT domain, also the consumers themselves may have different requirements in terms of how the data has to be consumed[5]. While this is ok in push based ICN architectures like Mobilityfirst[6], it s a challenge in CCN/NDN[7] architectures which is designed for PULL based applications. Inter-connecting numerous IoT entities, as well as establishing reachability to them, requires a scalable name resolution system considering several dynamic factors like mobility of end points, service replication, in-network caching, failure or migration. The objective is to achieve scalable name resolution handling static and dynamic ICN entities with low complexity and control overhead.

Scalability: Scalability has to be addressed at multiple levels of the IoT architecture spanning naming, security, name resolution, routing and forwarding level. In ICN-IoT, the name resolution is performed at the network layer, distributed within the entire network. Thus, it can achieve high degree of scalability exploiting features like content locality, local computing, and multicasting.

---

[29] Claudio Compolo, Daniel Corujo et al “Information-centric Networking for Internet-of-things”, *IEEE Networks*, Jan , 2015.

[4]Baccelli, E. et al, "Information Centric Networking in the IoT:Experiments with NDN in the Wild", *ACM-ICN SIGCOMM* 2014

[5] J. Quevedo et al, “Consumer driven Information Freshness Approach for Content Centric Networking”, IEEE, NOM, 2014

[6] NSF FIA project, MobilityFirst., "http://www.nets-fia.net/", 2010.

[7]Van Jacobson et al, “Networking Named Content”, ACM, CoNext, 2009

Research Gap: ICN scalability is another area of active research not only in the web content space where one has to deal with billions of named content objects, but also in the IoT space where even the number physical assets could be orders or magnitude greater. Further scalability is demanded from the data generated by these devices. In addition a unified ICN-IoT framework has to deal with mobility of end points, ad-hoc communication, migration of services and dynamic evolution of content in the network. ICN is promising candidate to meet these challenges because of its inherent ability to distribute intelligence leveraging name-based networking, contextualized communication (e.g. scopes), caching , computing, and trust models to the level of D2D communications without the unnecessary overhead of imposing centralized client-server communication models which IP suffers from today. Scalability in IoT can also be handled considering hierarchical deployment model of IoT applications where large number of named entities and thus data generated will have only local significance without requiring global reacheability.

Resource efficiency: IoT devices can be broadly classified into two groups: resource-sufficient and resource-constrained. In general, there are the following types of resources: power, computing, storage, bandwidth, and user interface. In ICN-IoT, in both the constrained and non- constrained parts of the network, light weight ICN stack over L2[8], along with features such as in-network processing allows only data that are subscribed by applications in the specified context to be delivered. Thus, it offers a resource-efficient solution.

[8] Baccelli, E. et al, "Information Centric Networking in the IoT:Experiments with NDN in the Wild", *ACM-ICN SIGCOMM* 2014

-

Research Gap: The challenge here is to develop light weight ICN network layer and associated middleware[9] to ensure long battery life while being aware of computation and memory limitations of embedded systems. Also in general the overall initiative to make networking power efficient[10] requires distributing computing intelligence so that data can be filtered at various points saving significant router processing overhead otherwise spend on raw content transfer to the cloud as done in IP today. Further another mode of achieving resource efficiency is leveraging flexible deployment options of ICN as an end-to-end protocol, or applying well known resource efficient solutions in the constrained segments, and applying more relatively complex ICN stacks in the infrastructure.

Traffic Pattern. IoT traffic can be classified as local or wide area network scope depending on the network context. Local traffic generation is due to D2D interactions or device-to-infrastructure interaction during data push or pull operations. Wide area traffic contribution either through need for distributing IoT data or interaction of end points with IoT services. In ICN-IoT, one can easily cache data or services in the network, hence making more communications within local distances and reducing the overall traffic volume, for example all the sensors and actuators in a building management system (BMS) could self-organize, with minimal control plane support and without manual configuration of each device.

Research GapPredicted traffic pattern in ICN is a projection of traffic patterns of known IP applications today. In that sense, general ICN traffic pattern will only be known when there is wide deployment of ICN networks and applications, which will also include applications unknown today. Even considering today's traffic pattern, significant communication can be localized without any need for control plane infrastructure, for e.g. all the sensors and actuators in a building management system (BMS) could self-organize, with minimal control plane support and without manual

configuration of each device. The research challenge here to understand these traffic patterns and building scalable ICN-IoT middleware protocols that can self-organize in a local context, requiring manual intervention at designated points in the network where policy restrictions and exposure of the content or services are required to the external world.

Context-aware communications. . ICN exposes several contexts to the network beginning with names to allow efficient inter-connection between the consumers and producers; this is in contrast to transport networks like IP/MPLS where the network is not expected to use any application context for its own benefit. Beyond names, many IoT applications shall rely on contextual information such as social, relationships of owners, administrative groupings, location, type of ecosystem (home, grid, transport etc.) of devices and data (which are referred to as contexts in this document) to initiate dynamic relationship and communication. ICN-IoT supports contexts at different layers, including device layer, networking layer and layer. Contexts at the device layer include information such as battery level or location; contexts at the layer include information such as network address and link quality; contexts at the application layer are usually defined by individual applications. In ICN-IoT, device and network layer contexts are stored within the network, while network elements (i.e., routers) are able to resolve application-layer contexts to lower-layer contexts. As a result of adopting contexts and context-aware communications, communications only occur under certain contexts that are specified by applications, which can significantly reduce the entire network traffic volume.

Research Gap: Contextualized communication in ICN-IoT is a powerful feature which adds another dimension of intelligence to the overall name-based communication model. For contexts to be understood at the network layer, in-network computing has to be the key enabler. While it will not be feasible to process every context relevant to all services in the network layer, abstraction of contexts that a wide majority of applications can benefit from has to be supported. In this context, efficient forwarder implementation with flexibility to process “well-understood” contexts and also able to customize in-network processing through service “plug-ins” is desirable; a good example is to enable big-data analytics only certain points in the network over certain stream of content flowing through the forwarder. This area in general is an open area of research[10,11] and will evolve as ICN applications mature and are more widely understood. In addition in the area of ICN-IoT security in the context of in-network processed data is another key consideration as data privacy and regulation is of at most importance.

[10] Y. Chen, A. Li and X. Yang, "Packet Cloud: Hosting In-Network Services in a Cloud-Like Environment," *Duke CS-TR-2011-10*.

[11]M. Sifalakis, B. Kohler, C. Scherb, and C. Tschudin: An Information Centric Network for Computing the Distribution of Computations, *1st International ACM Conference in Information Centric Networking (ACM ICN 2014)*, September 2014, Paris, France.

Seamless mobility handling. Mobility in the IoT platform can mean 1) the data producer mobility (i.e., location change), 2) the data consumer mobility, 3) IoT Network mobility (e.g., a body-area network in motion as a person is walking); and 4) disconnection between the data source an destination pair (e.g., due to unreliable wireless links). The requirement on mobility support is to be able to deliver IoT data below an application's acceptable delay constraint in all of the above cases, and if necessary to negotiate different connectivity or security constraints specific to each mobile

context. In ICN-IoT, ICN's name resolution layer should aid multiple levels of mobility relying on receiver-oriented nature for self-recovery for consumers, to multicasting and late-binding techniques to realize seamless mobility support of producing nodes.

Research Gap: Today mobility is restricted to mobile smart devices (phones, tablet etc.), but in a unified ICN-IoT platform that includes transport vehicles, planes, ships etc. enabled with many sensors and real-time reachability to these entities is critical at all times. Though ICN provides a good abstraction to applications, as they bind to persistent IDs, ICN layer has to scale to accommodate resolution of billions of mobile devices while meeting stringent application real-time requirements, which includes 1-10ms delay requirements required in the IMT-2020 architecture. Many distributed mobility techniques are being studied[12,13,14,15] to handle named-entity mobility which should also address the challenges of ICN-IoT applications as well. Also mobility handling varies with specific ICN protocol fundamental design objectives[4], hence choices have to be carefully made while enabling new features such as mobility in protocols like CCN.

[12] Azgin, A., Ravindran, R., and G. Wang, "A Scalable Mobility-Centric Architecture for Named Data Networking.", *ICCCN (Scene Workshop)* , 2014.

[13] Zhang, Y., Zhang, H., and L. Zhang, "Kite: A Mobility Support Scheme for NDN.", *NDN Technical Report NDN-0020* , 2014.

[14] Jordan Auge, Giovanna Carofiglio et al, "Anchor-less Producer Mobility in ICN*", ACM-ICN, SIGCOMM*, 2015.

[15] Li, S., Zhang, Y., Dipankar, R., and R. Ravindran, "A comparative study of MobilityFirst and NDN based ICN-IoT ", *IEEE, QShine*, 2014.

Caching and Storage: Storage and caching plays a very significant role depending on the type of IoT ecosystem, also a function subjected to privacy and security guidelines. In ICN-IoT, data are stored locally, either by the mobile device or by the gateway nodes or at service points. Also in-network storage/caching also speeds up data delivery.

Research Gap: While ICN's distributed short term buffers allow immediate content distribution of sensed and processed data and speed up data delivery[16] , long term storage can be used to drive analytics to allow efficient business processes. Also caching can be very useful in an ad hoc scenario, such as V2V[17] to provide DTN capabilities to end applications. In general the challenges here are to allow application-centric caching techniques considering application priority, QoS requirement such as latency or recovery due to mobility. Specific challenges apply to a given IoT scenario where caching feature has to account for security requirements such as access control and privacy. Caching has also been shown to be beneficial in constrained scenarios[18,] but its overall usefulness considering data privacy and access control in constrained environment has to be further studied. Another consideration toward caching is conflicting requirement by applications to have access to the latest data or data within a certain specified freshness while the name of the latest data by the producer is unknown by the consumer[19.] ICN is well suited to handle short term congestion, transport over unreliable links, and long term delay tolerant communications challenges using the caching/storage feature in the forwarders over every hop, providing reliable communications over unreliable links.

[16] Dong, L., Zhang, Y., and D. Raychaudhuri, "Enhance Content Broadcast Efficiency in Routers with Integrated Caching.", *Proceedings of the IEEE Symposium on Computers and Communications (ISCC)* , 2011.

[17] **Lucas Wang et al, "Rapid traffic information dissemination using named data",** *ACM workshop on Emerging Name-oriented Mobile Networking design*, 2012

[18] Baccelli, E. et al, "Information Centric Networking in the IoT:Experiments with NDN in the Wild", *ACM-ICN SIGCOMM* 2014.

[19] J. Quevedo et al, "Consumer driven Information Freshness Approach for CCN",*IEEE NOM*, 2014

Security and privacy. IoT security spans trust management challenges, security challenges includes data integrity, authentication, and access control at different layers of the IoT platform. Privacy means that both the content and the context around IoT data need to be protected. These requirements will be driven by various stake holders such as industry, government, consumers etc. In ICN-IoT, secure binding between application-centric names and content instead of IP addresses to identify devices/data/services, is inherently more secure than the traditional IP paradigm [20].

Research Gap: Generally ICN's object flexible security and trust models with intelligent network level security enablement operating on name semantics and other metadata such as keys should satisfy several ICN-IoT security functions related to authentication, integrity, access control and privacy. Because of senstitivity of IoT data and physical assets, security and privacy should be addressed in an end-to-end manner spanning control/forwarding/service plane interactions spanning functions such as naming, in-network computing, name-resolution, routing, caching, and ICN-APIs. The ability to cache content and apply in-network services is directly related to the security and trust policies applied in the application layer, hence to enable ICN-IoT applications to operate over an ICN network while satisfying application requirements is a challenge. Further in constraint networks, where computing capacity is minimal light weight security and privacy implementations are required, or other lower layer mechanism can be used where ICN level security is not possible. These aspects are being studied in specific context like building management systems (BMS)[21] and Healthcare[22], but require a more broader study scope if multiple systems is to share the same ICN infrastructure.

[20] Nikander, P., Gurtov, A., and T. Henderson, "Host identity protocol (HIP): Connectivity, mobility, multi-homing, security, and privacy over IPv4 and IPv6 networks", IEEE Communications Surveys and Tutorials, pp: 186-204, 2010.

[21] Wentao Shang et al, "Securing Building Management Systems using NDN", IEEE, Network, 2014

[22] Jeff Burke, "NDN Network Environment: Open mHealth"

Communication reliability. IoT applications can be broadly categorized into mission critical and non-mission critical. For mission critical applications, reliable communication is one of the most important features as these applications have strong QoS requirements. Reliable communication requires the following capabilities for the underlying system: (1) seamless mobility support in the face of extreme disruptions (DTN); (2) efficient routing in the presence of intermittent disconnection, (3) QoS aware routing, (4) support for redundancy at all levels of asystem (device,

service, network, storage etc.). ICN-IoT supports delay tolerant communications, which in turn provides reliable communications over unreliable links.

Research Gap: ICN offers many ingredients for reliable communication such as multi-home interest anycast over heterogenous interfaces, caching, forwarding intelligence for multi-path routing leveraging state based forwarding in protocols like CCN/NDN. However these features have not been analyzed from QoS perspective when heterogenous types of traffic is mixed in a router, in general QoS for ICN is an open area of research. In-network reliability also come at the cost of a complex network layer; hence the research challenges here is to build redundancy and reliability in the network layer to handle wide range of disruption scenarios such as congestion, short or long term disconnection, or last mile wireless impairments being aware of forwarding performance tradeoffs and network layer complexity. Also ICN network should allow features such as opportunistic store and forward mechanism to be enabled only at certain points in the network, as these offer control and forwarding plane overhead affecting application throughput.

Ad hoc and infrastructure mode. Depending upon whether there is communication infrastructure, an IoT system can operate either in ad-hoc or infrastructure mode. ICN-IoT supports both applications operating in ad-hoc and infrastructure modes.

Research Gap: The challenge here is to realize a single ICN protocol which can operate in both modes simultaneously, for e.g. a vehicle should be able to communicate with other vehicles in an ad hoc manner, and be able to communicate with a wireless WAN infrastructure or RSU without introducing any ICN gateways. This is quite realizable as information-centric nature of communication which naturally allows unicast/multicast/broadcast modes of communication between consumers and producers of information departing away from fixed session based approach between hosts in IP.

-

Self Organization : The unified IoT platform should be able to self-organize to meet various application requirements, especially the capability to quickly discover heterogeneous and relevant (local or global) devices/data/services based on the context. This discovery can be achieved through an efficient platform-wide publish-subscribe service [22], or through private community grouping/clustering based upon trust and other security requirements. Here scope-based self-organization is required to ensure logical isolation between the IoT subsystems, which should be enabled at different levels device/service discovery [23,24] naming, topology construction, routing over logical ICN topologies, and caching which in general is an open area of research[25]

[22] Jiachen, C., Mayutan, A., Lei, J., Xiaoming, Fu., and KK. Ramakrishnan, "COPSS: An efficient content oriented publish/subscribe system", ACM/IEEE ANCS, 2011.

[23] Ravindran, R., Biswas, T., Zhang, X., Chakrabort, A., and G. Wang, "Information-centric Networking based Homenet",IEEE/IFIP, 2013.

[24] C. Westphal, B. Mathieu, O. Amin, A Bloom Filter Approach for Scalalbe CCN-based Discovery of Missing Physical Objects, invited paper IEEE CCNC'16, Las Vegas, January 2016

[25] Li, S., Zhang, R. Ravindran , Lijun Dong, Qinji Zhang, Y., Dipankar, R., and, G.Q.Wang, "IoT Middleware Architecture for Information-Centric Networking", Globecomm, ICN Workshop, 2015.

Research Gap: General IoT deployments involves heterogeneous IoT systems or subsystems within a particular scenario co-existing on the same wireless or wired infrastructure. Here scope-based self-organization is required to ensure logical isolation between the IoTsubsystems, which should be enabled at different levels device/service discovery, naming, topology construction, routing over logical ICN topologies, caching which in general is an open area of research. These challenges extended to constrained devices should be energy and device capability aware. In the infrastructure intelligent name-based routing, caching, in-network computing techniques should be studied to meet scope-based self-X needs of ICN-IoT.

In-network processing. IoT requires data processing at multiple levels ranging from PAN/LAN/WAN etc, and this requires seamless placement of services and its discovery. In-network computing enables ICN routers to host heterogenous services catering to various network functions and applications needs. Contextual services for IoT networks require in-network computing, in which each sensor node or ICN router implements context reasoning.

Research Gap: In-network computing increases the scalability and efficiency of the system, and with ICN name-based anycast allows to place services any where in the network. This flexibility also allows dynamic service placement based demand for various content and services. Generally this requires the network infrastructure to support computing feature as an integral design of the network layer, and this should be designed to not affect the forwarding capacity of the forwarder itself. For a WAN scale ICN-IoT deployment, challenges related to service orchestration, resource management, QoS to meet various ICN-IoT requirements have to be understood.

- Content Distribution
- Multicast (pseudo and real-time)
- …

8 Relationship/Coexistence with Mobile edge computing

8.1 ICN Mobile Edge Computing

ICN is a natural platform to deliver edge services. Its ability to host services, content, with other features such as in-network processing, caching/storage, and multicasting/mobility allows distributed service intelligence and large scale content distribution capability. A distinguishing feature of ICN based edge service platform is contextualized service delivery. Such a platform will manifest the idea of pushing the frontier of computing services and applications away from centralized nodes to the logical extremes of the network. These edge-service realizations can be located in the vicinity of the operator's Central Office (CO) or at Points-of-Presence (PoP), or eNodeB, or all the way into specific application context such as home networks or transport infrastructure enabling local instantiation of ICN services to consumers, while providing global service delivery through a unified ICN based service control infrastructure. Following points motivates this realization:

ICN-MEC– ICN with SDN/NFV integration:

Though ICN is an ideal platform for efficient application delivery considering its features to adapt to network disruptions and temporal evolution of services and content, these features also make

ICN a complex protocol if it were to be deployed to handle all network functions through distributed control plane mechanisms. Instead ICN should explore design choices offered by NFV/SDN [26] to handle network services like mobility or name resolution functions within practical limits of scalability, which is mostly well handled within local domain scenarios. Further these core ICN functions can be application driven paving way for ICN based service virtualization. Even functions like service chaining[27] is more a information-centric operation than host-centric which can be handled by ICN, even in the context of IP flows. Even more service agility via service-chaining can be realized over ICN transport with distributed service functions and dedicated application controllers to aid Interest or Data processing through policy based paths determined by the network services. The co-existance of ICN with SDN/NFV and realizing information-centric service virtualization with rich contextualized content delivery in general is an open research topic.

Research Gap: Includes the areas of using base SDN/NFV for ICN slicing while realizing SDN-ICN, NFV-ICN equivalent service planes. Realize ICN Centric Service Virtualization (compute, bandwidth, cache, storage) over NFV-ICN. Realize ICN Centric Network Virtualization [28,29] logical/physical separation of ICN forwarder resources for heterogeneous ICN flows) over SDN-ICN

[26] Ravi Ravindran et al, "Towards Software Defined ICN Based Edge Cloud Services", *IEEE, CloudNet*, 2013

[27] Mayuthan Arumaithurai et al, "Exploiting ICN for Flexible Management of SDN", ICN, Siggcomm, 2014

[28] A. Chanda, C. Westphal, D. Raychaudhuri, Content Based Traffic Engineering in Software Defined Information Centric Networks, in IEEE INFOCOM Workshop NOMEN'13, April, 2013. pdf

[29] A. Chanda, C. Westphal, ContentFlow: Mapping Content to Flows in Software Defined Networks, in Proc. IEEE Globecom, December 2013 (arXiv preprint arXiv:1302.1493, January 2013). pdf

ICN-MEC – Service Contextualization: : Services are best delivered by locally customizing them to what users want, because users who are located in different locations have different needs and requirements based on their context. Context can be defined as any information that is used to describe the state of an entity and can be classified as being user-, network-, service-, device- centric. Modeling and reasoning of heterogeneous contextual information involves a trade off between complexity of reasoning and expressiveness of data. Mapping the contextual information into service level and network level requirements leads to the challenge of federated and standardized semantic representation of state and context. On the other hand, users themselves access services through heterogeneous devices, network connectivity, with subjective preferences. Moreover, mobility considerations require services to be delivered from the best vantage point. An edge service framework that can handle these dynamic requirements with minimum overhead is desirable. Context of users and services can be inferred and semantics (meaning of data items within a context ontology) of content predicates can be interpreted towards an intelligent dissemination of services and content items. In addition to a physical view of network topology, ICN based services and applications can benefit from a higher level perspective of topology.

Research Gap: Generally, ICN research has been limited to study fundamental networking issues related to congestion control, name resolution, multicasting, caching using name-based networking. In-network computing is another ICN dimension which can fundamentally distinguish it from IP networking. This is so because, ICN deals with content objects with metadata associated for security, temporal descriptions, application-centric metadata, which can be correlated with consumer's intent [30, 32]; and conduct immediate transformation if the intent doesn't satisfy the available content. Modeling and reasoning of heterogeneous contextual information involves a trade off between complexity of reasoning and expressiveness of data[31]. However the data transformation also has security implications considering the content object based security model. Integration of generic and application-centric in-network functions towards service contextualitzation to aid security function offloading, transcoding, data aggregation and filtering is opegenerally an open area of research.

[30] P. Talebifard, R. Ravindran et al, "An Information Centric Networking Approach Towards Contextualized Edge Service ", *IEEE, CCNC,* 2015

[31] P. Talebifard and V. Leung, "A Dynamic Context-Aware Access Network Selection for Handover in Heterogeneous Network Environments", *Proceedings IEEE INFOCOM Workshops*, April 10-15, 2011 in Shanghai, P.R. China., pages 385–390.

[32] F. Bronzino, S. Stojadinovic, C. Westphal, D. Raychaudhuri, Exploiting Network Awareness to Enhance DASH over Wireless, IEEE CCNC, Las Vegas, January 2016

***MEC Service APIs*:** ICN realizes a new transport using information-centric APIs, this directly elevates the operator as an information-pipe rather than bit pipe provider. Operating an ICN platform, makes the operator an active intermediary to satisfy information requests from consumers. Further service virtualization features can be built over it using control and data plane to handle name resolution, mobility, multicasting etc. From adoption perspective, ICN platform should offer significant technical and business benefits through rich Service APIs to the operatorto services compared to current existing infrastructure considering the the new CAPEX and OPEX in investments towards deploying ICN, which is are still a topic of open

Research Gaps: Topics of research include Service-API for ICN deployment leveraging NFV/SDN framework, SLA definitions for Services over an ICN platform

8.2 *ICN Based Edge Service Framework*:

<TBD>

9 Relationship/Coexistence with Network Slicing

9.1 Use of ICN for Inter-function transport

In the future IMT-2020 network, where a high degree of network function virtualization is expected either explicitly in NFV or in network slices, the use of an ICN protocol for inter-function communications is an attractive option because it decouples the location of services from the

service request. The concept of using ICN in NFV and SDN is studied in several academic works[30][31][32][33]. ICN is well suited for service oriented routing, because each element in a service chain can name the next element in the service chain without the need for an external name resolver or manual configuration.

The initial deployment of virtualized functions inside a network Slice could use ICN technologies immediately, as these are green-field services realized within a provider network[34]. Initial implementations may require running ICN as a transport protocol over IP due to initial lack of hardware support for native ICN transport, but this would be a transitional situation. Even when running over an IP network layer, ICN services can still provide robust communications and endpoint discovery using common existing IP techniques (e.g. mDNS, DNS SRV records, and distribute rendezvous techniques, among others).

As data-plane programmable equipment enters the marketplace, native ICN slices and transport can offer high-performance ICN/CCNx services. For example, abstracted 5G services in a CCNx slice, such as MME, S-GW, and P-GW, could be implemented in software, in software with hardware assist, or in deep-programmable hardware. In each case, the abstracted Slice service looks the same, though each would offer different performance curves in terms of latency, throughput, and capacity. Likewise, for inter-NFV communications, native wire-speed equipment, which is being demonstrated in 2015 by some manufacturers, would improve throughput and flexibility compared to using an IP-overlay, but should not limit such deployments within the 5G timeframe.

The advantage of using ICN as the inter-function transport protocol, even in a transitional period over IP, is that it positions carriers to move to an ICN native approach, which can discovery, recover, and utilize carrier networks efficiently without centralized bottlenecks or points of failure. ICN approaches also position the carrier for dynamic service mobility and deployment without needing to track an IP underlay.

---

[30] Latre, Steven, et al. "The fluid internet: service-centric management of a virtualized future internet." *Communications Magazine, IEEE* 52.1 (2014): 140-148.

[31] Ravindran, Ravishankar, et al. "Towards software defined ICN based edge-cloud services." *Cloud Networking (CloudNet), 2013 IEEE 2nd International Conference on*. IEEE, 2013.

[32] TalebiFard, Peyman, et al. "Towards a context adaptive ICN-based service centric framework." *Heterogeneous Networking for Quality, Reliability, Security and Robustness (QShine), 2014 10th International Conference on*. IEEE, 2014.

[33] Nakao, Akihiro. "Software-defined data plane enhancing SDN and NFV." *IEICE Transactions on Communications* 98.1 (2015): 12-19.

[34] Nakao, Akihiro. "Application Specific Slicing for MVNO through Software-Defined Data Plane Enhancing SDN." *IEICE Transactions on Communications* ????.

9.2 Use of ICN function state migration

ICN is a good technology choice for function state migration and execution context migration in future 5G networks . Content Centric Networking (CCNx) facilitates process migration while enabling many desirable features such as strong checkpointing and data de-duplication. Not all migration techniques require strong checkpointing, and in those cases CCNx offers a faster and weaker naming technique that allows pages or blocks to go dirty in a checkpoint.

CCNx offers an intuitive naming of resources that are part of a service or execution context migration and building checkpoints around those resources for consistent state transfer.

De-duplication is a technique where only one copy of data exists and it is shared between multiple instances. CCNx allows resources to be de-duplicated both within and between virtual machine instances. For example, in the previous discussion about using hash names for resources, if two disk blocks, for example, have the same hash value they will refer to the same Content Object. Only the block index in the CCNx manifest will be different.

A VM hypervisor may also share blocks between VMs. When generating the names used to fetch a checkpoint, the source migration agent running in the source hypervisor could use a name like {/nyc/host7, hash = 0x63223…} so any instance or any component can share the same data. Assume that the memory page size and the disk block size are the same. Then that name for hash 0x63223… could be both a disk block and a RAM page of the same data (e.g. a shared library code section). Because the manifest can point to different name prefixes for each hash and can indicate the virtual resource of that hash, we can have the same physical bytes used for many purposes. This approach may also be applied when page and block sizes are not the same by using smaller units of naming.

9.3 Use of ICN for de-abstracting the network

In a traditional IP-based network, the Internet Protocol adds a level of abstraction to communications endpoints by assigning them a location-dependent name. When applications wish to communicate, they must rendezvous those host addresses – such as with DNS or SIP or some other well-known means – before communication can take place. Because the rendezvous is done outside the network layer, the rendezvous protocol must employ its own means to determine locality – such as with ping triangulation – to determine which replica to use. Some applications use IP anycast addressing to move rendezvous back in to the network layer and realize those benefits. CCNx, as an ICN protocol, naturally keeps rendezvous on addresses within the network layer so all applications can benefit from localized services without needing to add on additional rendezvous layers with their own localization protocols.

ICN may be used as a de-abstraction layer for virtualized functions: using direct function naming in ICN means the network can move functions and change routing without needing to update intermediate abstractions of endpoint identification. For example, a single host IP address might hide many virtualized functions, so it may not be possible to directly move an IP address. One would need orchestration to inform components of a new socket endpoint, which could result in service interruption during the time when a function has finished migration and the time when an existing prior service is notified of a new service endpoint. With CCNx, the orchestration does not

need to inform prior components of a new service address, it only needs to update the named routing to the new location.

Because CCNx, as an example ICN protocol, is not tied to the P-GW identity – such as for the source endpoint address – it means that CCNx is well suited for multiple P-GW egress. Service frameworks, such as Mobile Edge Computing, could realize significant simplification by using a CCNx approach for multiple P-GW egress without needing to assign the UE multiple identities or using layers of address translation.

9.4 ICN gaps related to Sotwareization

Gaps

3. Use of ICN within a slice would require ICN-aware components in the Slice, such as specialized software or deep-programmable elements to execute the ICN protocol.
4. For ICN to be an effective rendezvous mechanism in the routing plane, the CRAN would need to use the ICN protocol.

10 Identification of gaps

The following ICN gaps were identified. See clause 7.5.1 of the main body of this report of the FG-IMT-2020 focus group for the detailed description.

Gap E.1: Considering ICN as a protocol for IMT-2020 Network

Gap E.2: Robust header compression for air interface (PDCP)

Gap E.3: Mobility anchoring (ICN aware S-GW)

Gap E.4: Mobility (ICN-aware MME)

Gap E.5: ICN Protocol (ICN-aware P-GW operation)

Gap E.6: ICN Protocol Execution (slice)

Gap E.7: Lawful intercept (specify what to capture)

Gap E.8: ICN mobility and routing

Gap E.9: ICN UE provisioning

Gap E.10: ICN managing IMT-2020 Self Organizing Network (SON)

Gap E.11: ICN – Operations and management (common interfaces)

Gap E12: Operations and management (SDN/Openflow)

Gap E.13: Security (authentication and encryption)

Gap E.14: Security (encryption)

Gap E.15: QoS (demand based)

11 Migration from the existing network technology

Editor's note: This clause is empty.