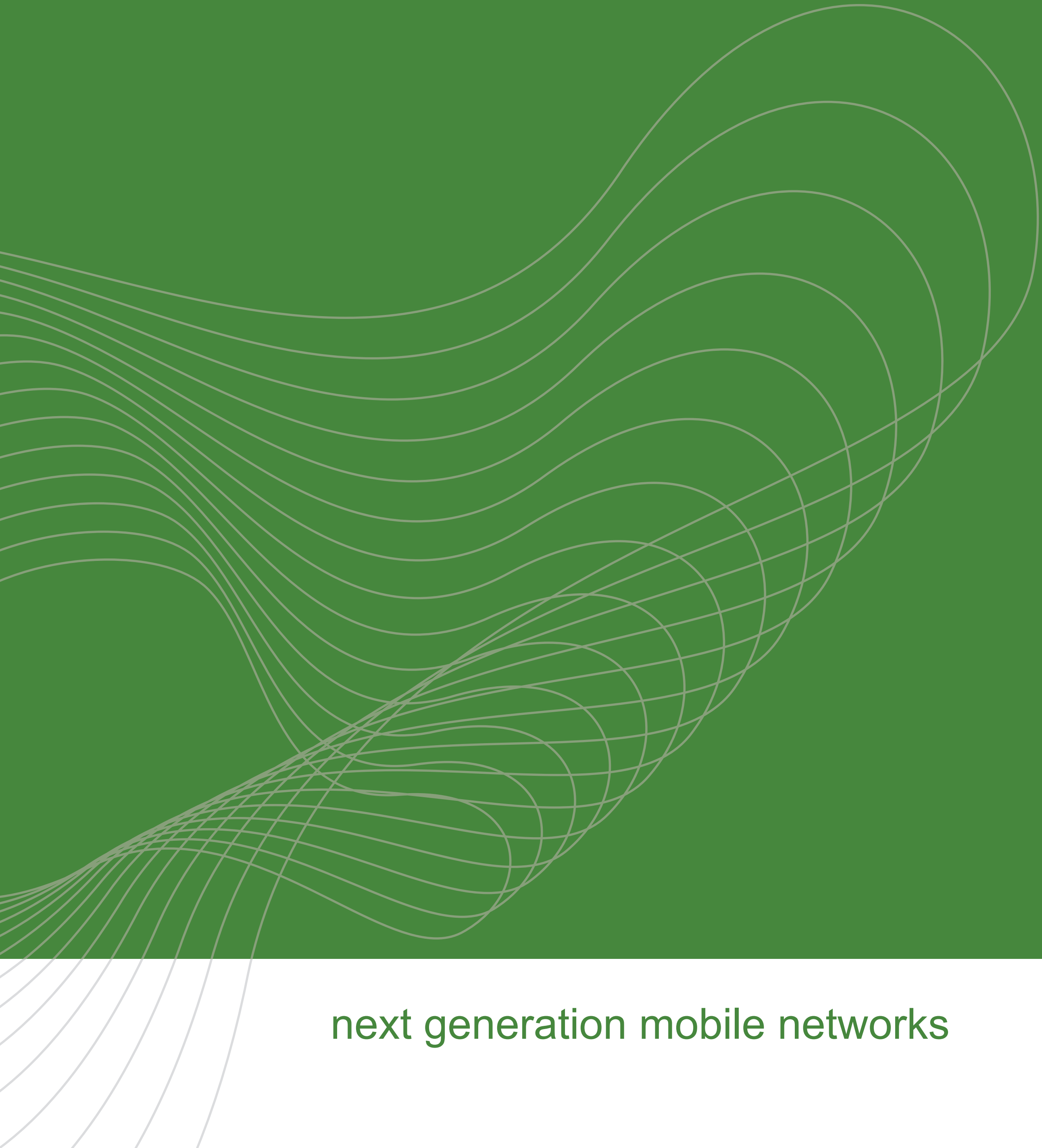
ngmn
the engine of broadband
wireless innovation

5G Extreme Requirements: End-to-End Considerations



next generation mobile networks

5G Extreme Requirements: E2E Considerations

by NGMN Alliance

Version:	**2.0**
Date:	**16-Aug-2018**
Document Type:	**Final Deliverable (approved)**
Confidentiality Class:	**P - Public**
Authorised Recipients: (for CR documents only)	

Project:	**5G Extreme Requirements Task Force**
Editors / Submitters:	**Ilaria Thibault, Vodafone; Sophie Vrzic, Huawei**
Contributors:	**Jaya Rao, Huawei; Joachim Sachs, Ericsson; Alan Stidwell, Orange; Hakan Batikhan, Turkcell; Kevin Smith, Vodafone; David Lister, Vodafone; Javan Erfanian, Bell Canada;**
Supporters:	**Alexander Chassaigne, Telefonica; Anass Benjebbour, NTT DoCoMo;**
Approved by / Date:	**NGMN Board, 6th August 2018**

Commercial Address:
ngmn Ltd.,
Großer Hasenpfad 30 • 60598 Frankfurt • Germany

Phone +49 69/9 07 49 98-04 • Fax +49 69/9 07 49 98-41

Registered Office:
ngmn Ltd.,
Reading Bridge House • George Street • Reading •
Berkshire RG1 8LS • UK
Company registered in England and Wales n. 5932387,
VAT Number: GB 918713901

Abstract:

The aim of this work is to highlight what implications and trade-offs related to the delivery of new 5G services are relevant for mobile network operators. Some of these new services, in fact, require extremely low latency and high reliability of the communication link, which have very little in common with the targets that the telecommunications industry has worked towards until today. The new 5G requirements, in fact, now call for a re-think on how the future network will have to be designed and optimised in order to enable the new services.

The purpose of this document is to identify realistic end-to-end deployment configurations that can potentially deliver the 5G extreme services across their footprint and to highlight some of the key challenges that come into play in this context.

Document History

Date	Version	Author	Changes
February 9th, 2018	0.0	Ilaria Thibault	Outline and Introduction
February 16th, 2018	0.1	Ilaria Thibault	Added, scope, definitions and assumptions
February 20th, 2018	0.2	Sophie Vrzic	Added Sections 6, 7, 8, 9
February 21st, 2018	0.3	Ilaria Thibault	Accepted changes and added comments.
February 23rd, 2018	0.4	Sophie Vrzic	
March 6th, 2018	0.5	Ilaria Thibault	
	0.6	Sophie Vrzic	
April 12th, 2018	0.7	Sophie Vrzic	
April 12th, 2018	0.8	Ilaria Thibault	Added Executive summary and conclusion
April 13th, 2018	0.9	Ilaria Thibault	
May 4th	1.0	Sophie Vrzic	Addressed all of Ericsson's comments in sections 6 and 7 and some in the conclusion section.
May 10th, 2018	1.1	Ilaria Thibault	Addressed remaining comments from Ericsson
May 22nd, 2018	1.2	Sophie Vrzic	
May 23rd, 2018	1.3	Ilaria Thibault	
May 24th, 2018	1.4	Ilaria Thibault	Clean version
June 8th, 2018	1.5	Sophie Vrzic, Ilaria Thibault	Addressed comments from Ericsson
June 21st, 2018	1.6	Sophie Vrzic	Addressed comment in Section 8.2 and outstanding comments in conclusion
July 3rd, 2018	1.7	Ilaria Thibault	Cleaned up version
July 11th, 2018	1.8	Sophie Vrzic	Addressed final comments from Kevin Smith, Vodafone, and fixed typos
July 26th, 2018	1.8-v2	Hans J. Einsiedler	Some typos and explanations
August 14th, 2018	1.9	Ilaria Thibault, Sophie Vrzic	Addressed final comments from DT and China Mobile
August 16th, 2018	2.0	Ilaria Thibault	Fixed references

Contents

1 EXECUTIVE SUMMARY

The next generation of mobile networks is currently being designed to deliver new services, which will enable new business opportunities in partnership with new vertical players.

The applications that rely on the 4G infrastructure today are able to cope, to a certain extent, with variations in data rate, reliability and latency, which naturally occur in response to varying channel conditions on the end-to-end signal path.

5th generation mobile networks will support a wide range of new services, with requirements that strongly deviate from the traditional mobile broadband targets. Some new services, denoted as Ultra-Reliable and Low-Latency, inherently cannot tolerate variations in data rate, reliability and latency, as the consequences of these variations could translate into the failure of critical infrastructure. Hence, the end-to-end network needs to be designed in a robust and reliable way so that wherever the user is in the service coverage area, it will experience the same guaranteed quality of service.

These new 5G use cases thus demand a paradigm shift in the packet transmission techniques applied both in the radio access network and in the core network to contend with the extreme requirements with complex trade-offs. For these target use cases it is exceptionally challenging and, resource usage-wise, prohibitively expensive to satisfy the extreme reliability and latency requirements using the existing architectures and protocols. Since the reliability requirement is intertwined with latency, both performance metrics have to be jointly considered in the overall end-to-end architecture and protocol design. This is because transmitting packets with high reliability is consequential only if the packets, traversing via multiple forwarding nodes and links, are received within the target latency bound.

From the deployment perspective, satisfying the extreme requirements on an end-to-end basis brings about new challenges because any of the considered techniques have to interwork across multiple domains, both within and outside the scope of 3GPP.

To this end, this work follows the studies on the radio access network published in [1] and [2], as it is pertinent to investigate and evaluate the different applicable techniques by comprehensively accounting for the latency and reliability at each node and link in the end-to-end transmission path. Based on the evaluations, the relevant network architectures, deployment techniques and transport protocols to support the target use cases can be determined.

The outcome of this deliverable can be summarised as follows:

- The factors that affect latency and reliability (separately) in the end-to-end communication path are analysed, both within and outside the 3GPP domain.

- A method for optimising latency and reliability jointly from an end-to-end perspective is proposed. The objective is to minimise deployment cost or allow for deployment flexibility of application servers and user equipment within a geographical area and, at the same time, to comply with end-to-end service level requirements on latency and reliability.

- A numerical analysis is provided to give guidelines on how to deploy an end-to-end system based on the considered optimisation goal. The end-to-end system relies on the specific radio access solutions identified in Phase 2.1 [2] and needs to meet latency and reliability service-level targets. The use cases studied in Phase 2.1 [2] and listed in Table 1 are used as a basis to carry out the analysis.

- Techniques such as path redundancy and new transport-layer protocols are discussed as a means to improve end-to-end latency and reliability.

Table 1 Targets for end-to-end numerical analysis

Target 1	Fast and reliable transfer of large messages.
Target 2	Fast and reliable transfer of small messages.
Target 3	Ultra-fast and ultra-reliable transfer of small messages.

2 INTRODUCTION AND MOTIVATION

New business opportunities for operators in a wide range of vertical industries (e.g., smart manufacturing, logistics, transportation, health, smart cities, agriculture, gaming, etc…) translate into new and sometimes challenging sets of targets that 5th-Generation mobile cellular networks need to meet to be able to successfully deliver the desired services. These targets include an evolution of traditional mobile broadband, which has been the main driver for network development until today, as well as requirements that are completely new to the cellular industry and that mainly address Internet-of-Things type of use cases, where, e.g., new industrial verticals may become customers.

In this context, a wide range of use cases with related business opportunities and required network capabilities was identified by NGMN in [3] and [4]. This work then became valuable input for 3GPP when it kicked off its own studies on new services for the next generation of mobile communications, summarized in [5]. 3GPP organised all the different use cases and their service-level requirements into three main categories: massive Internet of Things [6], Critical Communications [7], and enhanced Mobile Broadband [8]. These studies then formed the basis for a single specification [5]. In addition to these evaluations of service needs, ITU has defined the requirements for the 5G radio [9] [10], covering the same categories. Standardization work is now ongoing in different working groups within 3GPP [11] [12] aiming at meeting both these sets of requirements (service-level and radio link level). A refinement of use cases and requirements for vertical industries is currently ongoing in 3GPP [13] [14].

Massive-Internet-of-Things requires the network to support very large numbers of connections for machine-type traffic; Critical Communications often call for very low latency and highly reliable wireless access links for the delivery of advanced functionalities for controlling objects; and enhanced Mobile Broadband enables data-rich and immersive applications that rely on augmented and virtual reality features. Many actual use cases will extend into more than one category, and thus require enhancements from multiple dimensions such as coverage, quality-of-service, and capacity. To some extent, we can regard them as using different modes of the network; a long-range massive mode (massive Machine Type Communication mMTC), a highly reliable and low-delay mode (Ultra Reliable Low Latency Communication, URLLC), and a high-data-rate mode (enhanced Mobile Broadband, eMBB).

3 SCOPE

NGMN has recognised the need to gain deeper understanding in what impact these services will have on the future network architecture, both for the radio access and for the entire end-to-end network. Therefore, a task force on 5G Extreme Requirements was kicked off in May 2017. The new requirements are referred to as “extreme” since they go far beyond the boundaries of the traditional targets that have been the main driver for network design until today, and the focus of this work is the case when very high reliability and ultra-low latency are required at the same time.

This task force has the objective of answering the following questions:

1) To which extent can the 5G extreme services be delivered on existing deployments?
2) What modifications, if any, are required in the radio access network and/or in the core network to deliver the 5G extreme services?
3) How sensitive are the deployment models to the requirements? By relaxing the targets, does the deployment change considerably?

In order to answer the questions above, the 5G Extreme Requirements Task Force is structured into two main phases, which are mapped to a time line in Figure 1:

- **Phase 1: Operators' view on fundamental trade-offs**:
  This is a high-level study that provides preliminary insight for Question 1. The fundamental trade-offs among latency, reliability, message size, data rate, and service coverage area are analysed. More detailed and technology-specific analysis is the scope for Phase 2.

- **Phase 2: Network solutions for extreme services**:
  The objective is to identify how and to what extent the service-level requirements and the radio link requirements can be supported, and compare different end-to-end network solutions that address those critical sets of requirements. This phase aims at answering in detail Questions 1, 2 and 3 and is broken down into two sub-phases that address radio access and end-to-end aspects respectively, as described below.

  - **Phase 2.1: Radio Access Network solutions**:
    A given set of services associated with requirements on latency, reliability, throughput, and coverage availability is considered. Different Radio Access Network solutions are then applied, and their potential in being able to support the chosen services is assessed. Both LTE-Advanced (according to 3GPP Rel. 15) and New Radio (NR, defined in 3GPP Rel. 15) are considered as candidate radio access technologies with different bandwidth configurations.

  - **Phase 2.2: End-to-End considerations**:
    This phase extends the scope of Phase 2.1 by identifying what affects latency and reliability in an end-to-end deployment and which changes and new features are required from an end-to-end network perspective to meet the targets associated to the services identified in Phase 2.1.

This report outlines the outcome of Phase 2.2, which can be summarisedas follows:

- A framework is defined to outline end-to-end latency and reliability trade-offs, which highlights the fact that designing a network end-to-end means having to understand the complex interplay of the 3GPP system with non-3GPP solutions that provide higher layer functionalities, such as transport and application.

- A numerical analysis is carried out, which takes the results from Phase 2.1 as an input, and aims at dimensioning the rest of the end-to-end system based on service-level requirement targets. Numerical examples of maximum node and link failure rates, maximum service distance, required redundancy in the core and the the radio access network, density of edge-server nodes are provided as an output of the numerical study.

The report is structured as follows:

- Section 4 outlines key definitions, scenarios, and assumptions,
- Section 5 defines the target end-to-end requirements that are considered in this study,
- Section 6 describes the methodology adopted for the numerical analysis,
- Section 7 presents the results obtained through the numerical analysis,
- Section 8 presents considerations on tunnelling and protocol enhancements for extreme requirements,
- Section 9 presents considerations on how to satisfy extreme requirements in a mobile scenario,
- Section 10 concludes the work.

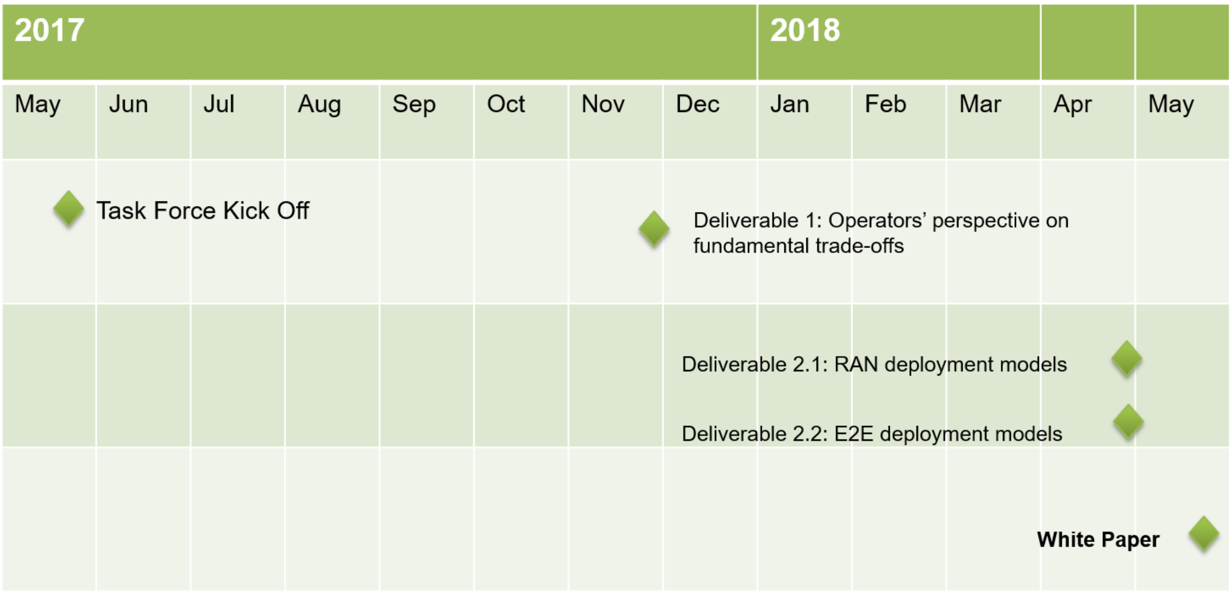


Figure 1 Extreme Requirements Task Force time line.

4 DEFINITIONS AND ASSUMPTIONS

In the context of this analysis, the term **end-to-end (E2E) network** is used. The two end points are an application running on an end system, and an application running on an **application server (AS)**. The **user equipment (UE)** is the communication module that allows the end system to communicate with the AS. In this context, we assume that the end system and the UE are collocated, and we refer to them as UE, for simplicity. In practice, a distribution of end systems, such as cameras, sensors, etc. could all interconnect to the same UE. E2E network denotes all the protocol layers, network functions, and infrastructure nodes that exist between the application layer on the end system and the application layer at the AS, in the context of a mobile network. The considered E2E network diagram is shown in Figure 2, where the Open System Interconnection model (OSI) is used for representing the protocol stack on each element of the communication chain. E2E denotes the signal path from the layer 6/7 interface on the end system to the layer 6/7 interface at the AS. The focus of this analysis is on the user plane, where the UE connects to the AS through an eNB/gNB, i.e., the radio access network (RAN), which in turn is connected to the User Plane Function (UPF), in line with 3GPP's Next Generation Core Network (NGCN) architecture [15]. We assume that the AS is co-located with the UPF, and that both entities are hosted by an edge node.

The term **Access Point (AP)** is used throughout the text to denote eNB or gNB.

Any type of infrastructure node placed between the AP and the AS is here called a **forwarding node**: this can be, e.g., a router, switch, point of concentration, repeater, etc.

The **E2E latency** is the time that takes to transfer application-layer data of a given size from a source to a destination, from the moment it is transmitted by the source to the moment it is successfully received at the destination (one-way latency). In other words, the E2E latency is measured from the L7/6 interface on the end system side to the L7/6 interface on the AS side, or vice versa.

The **E2E Reliability** is the percentage value of the amount of sent application layer packets successfully delivered to a given node within the time constraint required by the targeted service, divided by the total number of sent application layer packets. Measured from the L7/6 interface on the end system side to the L7/6 interface on the AS side, or vice versa.

It is important to note that the E2E system is a combination of the 3GPP system and a non-3GPP system, hence its performance is influenced by both these elements. The 3GPP domain, in fact, includes the signal path from the layer 2/3 interface on the UE to the layer 2/3 interface on the UPF, whereas, as defined above, the considered E2E network includes higher layers on the end system side, as well as all the layers on the AS side, as shown in Figure 2 and Figure 3. As represented in Figure 2 and outlined in Section 3, the scope of Phase 2.1 of this work has been to focus on the 3GPP radio access network performance, i.e., on the connection between the layer 2/3 interface on the UE to the layer 2/3 interface at the eNB [2].

The scope of Phase 2.2 is to take the Phase 2.1 results as an input, and provide guidelines on E2E network dimensioning based on service-level requirements. Figure 3 highlights the fact that E2E latency and reliability are influenced by the latency and reliability of both the 3GPP and non-3GPP systems.

The guiding principle of network dimensioning has been to maximise the physical distance between the UE and the AS, whilst delivering the required E2E latency and reliability for a given service. The distance between UE and AS is here defined as the **service distance**, and it is represented in Figure 4.

There are a few main reasons for choosing to maximise the service distance, subject to E2E latency and reliability requirements:

- This allows to minimise AS deployment costs, as the number of necessary edge nodes would be reduced to a minimum for a given set of requirements on latency and reliability.
- It provides a geographical boundary around the UE within which the AS needs to be deployed in order to satisfy the requirements. The operator can then flexibly decide where to deploy the edge node within this boundary.

Based on the chosen service distance, each edge node will serve a certain number of APs. The inverse of the number of APs served by a single edge node is here defined as **service density**, as represented in Figure 5. Maximising the service distance is equivalent to minimising the service density.

As mentioned above, in practice, when it comes to ultra-reliable and low latency services, a distribution of several real end devices, like video cameras, haptic sensors, etc. can all connected to the same UE. The analysis carried out in this work assumes that all processing layers are placed within a single processing entity, which we refer to as UE, for simplicity. However, the proposed methodology can also account for a distributed version of the end system if adequate parameters representing processing delays and reliability at different layers are chosen for the numerical evaluations. In this case, the service distance would need to be calculated from where the end device hosting the application is located.

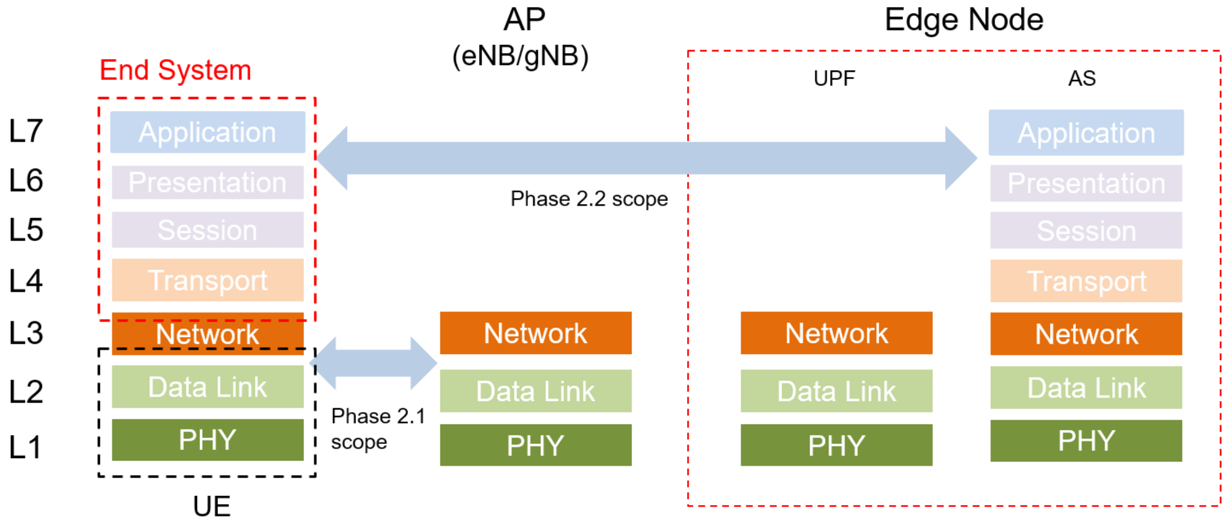


Figure 2 E2E network diagram, user plane. The following elements are represented (starting from the left of the diagram): End System, User Equipment (UE), Access Point (AP), Edge Node, which hosts the User Plane Function (UPF), and the Application Server (AS).

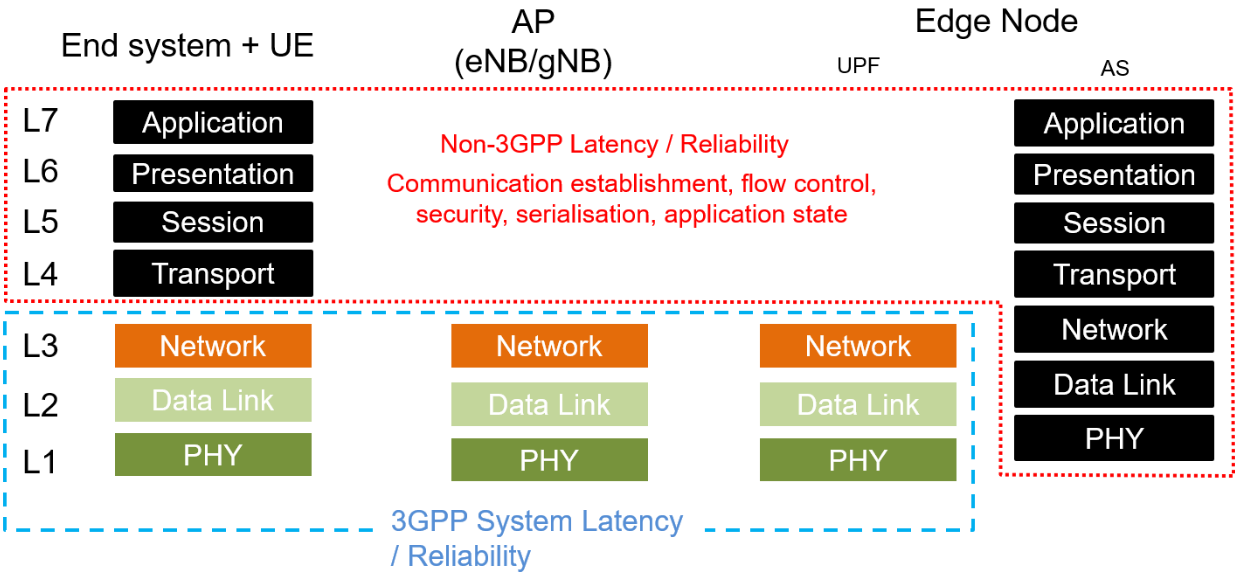


Figure 3 E2E Network diagram: latency and reliability depend on the performance of the combination of the 3GPP system and a non-3GPP system.

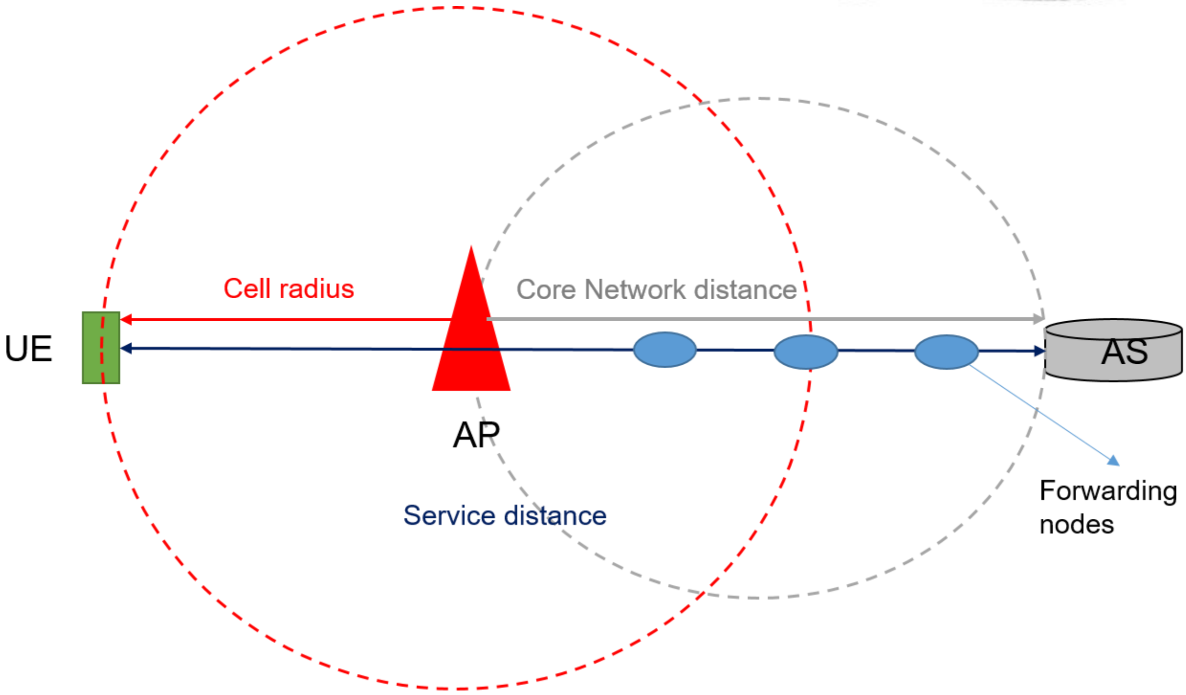


Figure 4 Service distance = Cell radius + Core Network distance

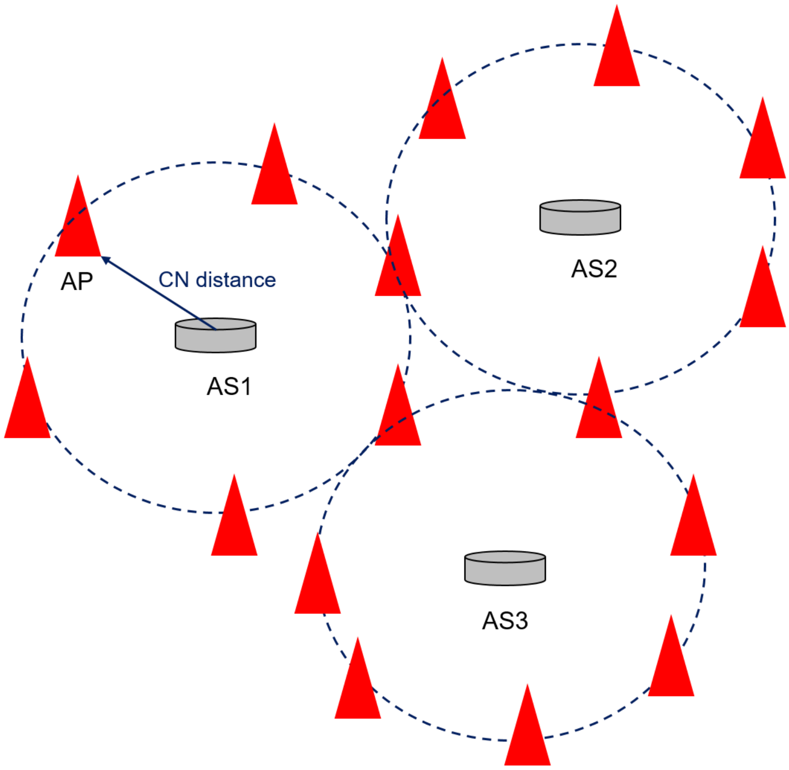


Figure 5 Service density (illustrative)

5 TARGET E2E REQUIREMENTS

The E2E requirements considered in this Phase 2.2 of the task force corresponding to the three use case categories identified in Phase 2.1 [2] are specified Table 2. For each use case, three different target E2E latency requirements are considered in order to analyse how sensitive the E2E deployment is to variations in service-level latency requirements [2].

Table 2: Use Cases.

Target	Examples of Use Cases	Payload	Target E2E Reliability	Target E2E Latency (ms)
1	Data-rich applications for media and entertainment such as, e.g., Augmented Reality, Virtual Reality, collaborative gaming, etc. This is a use case that requires fast and reliable transfers of large payloads.	1500 Bytes	99.9%	10.5
				11
				12
2	Small-payload applications for use cases that require interaction between sensors/actuators and a controller such as remote control for smart manufacturing, electricity distribution, etc… Typically the interaction happens in periodic patterns. Reliability is high as it reflects the need for robust wireless links.	40 Bytes	99.99%	5.5
				6
				7
3	This is the target with the most challenging requirements for 5G, and it represents use cases such as tactile interaction, discrete automation, etc… The payload requirement is low, but very high reliability needs to be met within an extremely low latency budget.	40 Bytes	99.999%	1.5
				2
				3

6 METHODOLOGY

6.1 Reliability Analysis

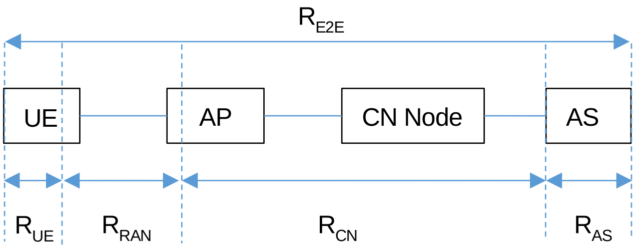
As defined in Section 4, the E2E reliability is the probability of correctly decoding an application layer packet at the receiver within the packet delay bound. Packets that arrive after the packet delay bound and packets that are lost or erroneous are considered as errors.

The E2E reliability depends on the reliability of the UE, RAN, CN and the AS. It can be represented by the following equation

$$R_{E2E} = R_{UE} R_{RAN} R_{CN} R_{AS} \quad (1)$$

where R_{UE}, R_{RAN}, R_{CN} and R_{AS} represent the reliability of the UE, RAN, CN and AS, respectively.

The elements that influence the E2E reliability are illustrated in Figure 6. The R_{RAN} is also referred to as the reliability on the access link (AL), R_{AL}.



Note: $R_{RAN} = R_{AL}$ from simulations in phase 2.1 [2]

Figure 6: End-to-end reliability

The reliability of the UE, R_{UE} and the reliability of the AS, R_{AS} include both hardware and software failures. For the purpose of this analysis, both R_{UE} and R_{AS} are assumed to be one, which corresponds to 100% reliability.

The reliability of the RAN depends on the probability of correctly decoding a packet within the packet latency bound, which depends on the probability of failure of the AL, denoted as $P_{f,AL}$. This value is the error rate that is determined in Phase 2.1 [2]. The RAN reliability is thus given by the following equation

$$R_{RAN} = R_{AL} = 1 - P_{f,AL} \quad (2)$$

The reliability of the CN, R_{CN} depends on the reliability of each link and node in the core network. It is given by the equation

$$R_{CN} = \left(1 - P_{f,link}\right)^{N_{links}} \left(1 - P_{f,node}\right)^{N_{nodes}} \quad (3)$$

$P_{f,link}$ is the probability of link failure,
$P_{f,node}$ is the probability of node failure,
N_{nodes} is the number of nodes in the CN,
N_{links} is the number of links in the CN

The number of nodes includes the AP, the edge node hosting the AS and the CN nodes in between the AP and edge node, such as switches, routers, points of concentration and UPFs. The CN nodes between the AP and edge node are referred to as **forwarding nodes**.

The total number of links and the total number of nodes in the E2E network can be expressed as a function of the total number of forwarding nodes that are placed between the AP and the edge server, N_i as given by the following equations.

$$N_{links} = N_i + 1 \qquad (4)$$
$$N_{nodes} = N_i + 2$$

The number of forwarding nodes is equivalent to the number of hops in the path from the AP to the edge node. This is illustrated in Figure 7.

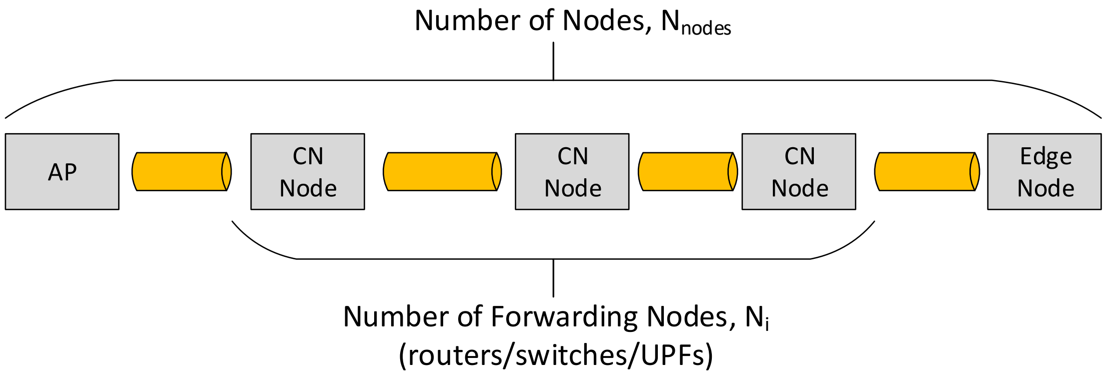

Number of Nodes, N_{nodes}
AP
CN Node
CN Node
CN Node
Edge Node
Number of Forwarding Nodes, N_i
(routers/switches/UPFs)

Figure 7: Forwarding Nodes in the CN.

The probability of link failure, $P_{f,link}$ depends on the transport medium (e.g., fibre, copper, microwave). Typical values for both $P_{f,link}$ and $P_{f,node}$ range from 10^{-6} to 10^{-4} [16] [17]. For simplicity, it is assumed that all of the nodes and links in the CN have the same probability of failure. However, the analysis can be easily generalised to the case where these probabilities have different values for each node and link.

Redundancy can be used to improve the reliability in the RAN and the CN. Adding redundancy translates into adding independent and parallel links to the existing ones, so that the signal can be replicated across multiple independent paths.

The reliability of the CN, after taking redundancy into account, is given by

$$R_{CN} = 1 - \prod_{i=1}^{N_{paths}} \left(1 - R_{CN}^{(i)}\right) \qquad (5)$$

where N_{paths} denotes the number of independent parallel paths. The term $R_{CN}^{(i)}$ represents the reliability of the path i in the CN. Each path in the CN can include multiple links through multiple forwarding nodes. This is illustrated in Figure 8.

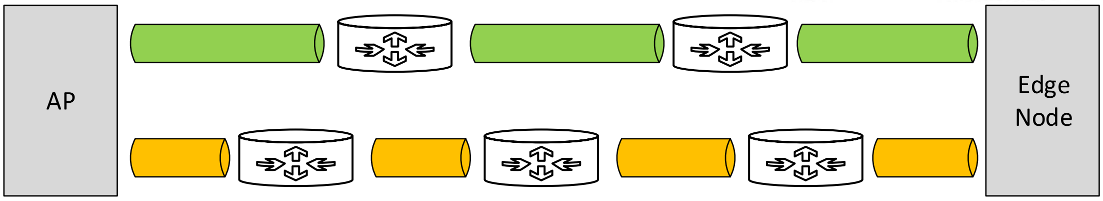


Figure 8: Redundancy Using Multiple Independent Paths.

In the above figure, each path between the AP and the edge node contains independent nodes and links. In this case, if there is a node or link failure on one of the paths then there is no impact to the other path.

Similarly, the reliability of the RAN can be improved using packet duplication, where the same packet is sent over two separate access links. For example, the UE can transmit/receive the same packet to/from two different access points. This can be represented by the equation

$$R_{RAN} = 1 - \left(1 - R_{AL}^{(1)}\right)\left(1 - R_{AL}^{(2)}\right) \tag{6}$$

where $R_{AL}^{(1)}$ and $R_{AL}^{(1)}$ is the reliability of the first access link and second access link, respectively.

The above equation assumes that the access links are independent. However, this is typically not the case since there is some dependence on a shared control plane. For simplicity, the results in this report assume the links are independent.

6.2 Latency Analysis

The E2E latency depends on the processing delays and transmission delays at the UE, RAN, CN and AS.

The E2E latency is given by the equation

$$t_{E2E} = t_{UE} + t_{RAN} + t_{CN} + t_{AS} \tag{7}$$

where each term represents the delays at UE, RAN, CN and AS, respectively.

The term t_{UE} includes the processing delay at the UE above the access protocol stack. This includes processing at the application, presentation, session, transport, and network layers. Similarly, the term t_{AS} includes the processing delays for the application, presentation, session, transport, and network layers at the AS.

The latency in the RAN, t_{RAN} includes the processing and transmission delays over the access link (i.e., from the layer 2/3 interface on the UE to the layer 2/3 interface on the AP) and depends on the radio access configuration adopted in Phase 2.1 [2].

There may be multiple forwarding nodes in the CN, which include routers, switches, points of concentration and UPFs. Each CN node contributes to the E2E delay. The latency in the CN is given by the equation

$$t_{CN} = t_{prop} + (N_i + 1)t_{Tx} + (N_i + 2)t_{process} + t_{TP} \tag{8}$$

where t_{prop} is the propagation delay that exists between the AP and the AS, t_{TX} represents the transmission delay for each link, N_i is the number of forwarding nodes (e.g., switches, routers, and points of concentration) in the CN, $t_{process}$ is the processing delay at each node (which is considered as the same for all nodes) and t_{TP} represents the latency associated with the tunnelling protocol between the AP and the UPF.

The propagation delay depends on the transport medium used for the CN links (e.g. fibre, copper or microwave).

The term t_{prop} can be represented using the equation

$$t_{prop} = \frac{d_{CN}}{v} \qquad (9)$$

where d_{CN} is the distance in the CN from the access point to the AS in the edge node and v is the velocity of the transmission over the selected medium (for simplicity we here consider $v = c = 3 \times 10^8$ m/s). For fibre links, the value $v = c = 2 \times 10^8$ can be used.

The transmission delay, t_{TX} depends on the packet size and on the transmission rate of the link and is given by

$$t_{Tx} = \frac{P_{size}}{\alpha C_{link}} \qquad (10)$$

where P_{size} is the packet size and α is the fraction of the link capacity, C_{link}, that is allocated for the target use case.

Figure 9 illustrates the processing delays at each node.

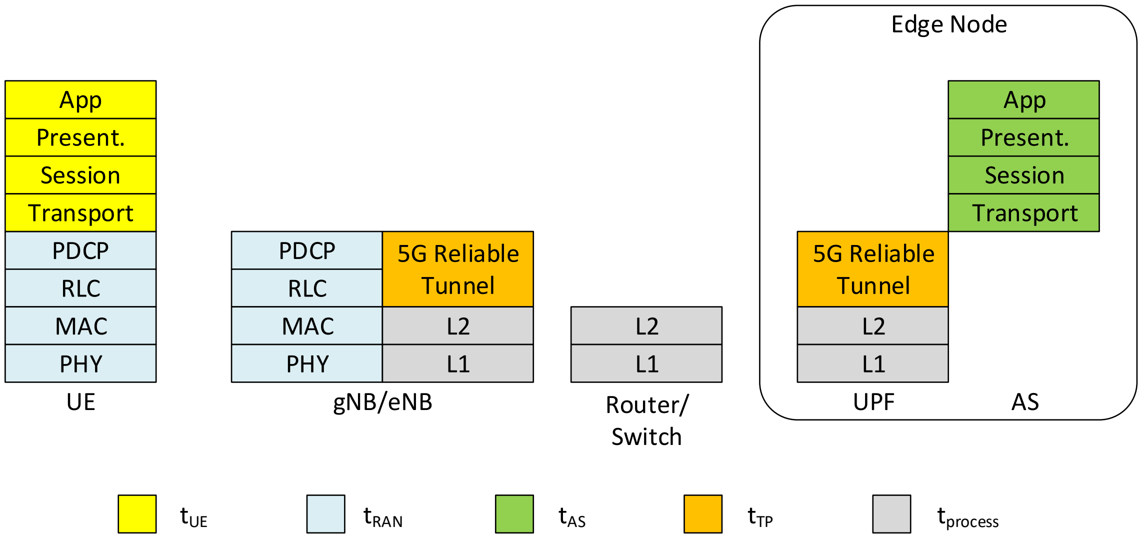


Figure 9: Processing Delays at the UE, RAN, CN and AS.

6.3 Joint Reliability and Latency Analysis

This analysis is based on the objective of maximising the service distance as defined in Section 4, which is the distance between UE and AS. The main motivation is to reduce the edge node deployment costs, or to allow for edge node deployment flexibility. A maximum service distance, in fact, characterises a specific boundary within which the E2E requirements can be met, which outlines an area within which the operator can flexibly deploy the service.

Since the signal transmitted by either the RAN node or edge node degrades over longer distances, forwarding nodes between the RAN node and the edge node, such as routers, switches, and points of concentration, can be used to increase the service distance. These forwarding nodes perform a relaying function by transmitting the received signal to the next node before signal degradation occurs.

However, this relaying function introduces an additional processing delay at each forwarding node. Therefore, there is a maximum number of forwarding nodes and maximum service distance that can satisfy the E2E reliability and latency requirements.

Since both the reliability and latency constraints depend on the number of links and nodes between the RAN nodes and the AS, there is a trade-off between the service distance and the E2E reliability and latency constraints. As the number of links and nodes in a single path increases, the reliability decreases, since there are additional points of failure. Similarly, as the number of nodes between the RAN nodes and the AS increases, the latency also increases, since there is an additional processing delay at each node.

The number of forwarding nodes impacts both the E2E reliability and the E2E latency. In order to satisfy both the reliability and latency requirements, the maximum number of forwarding nodes must be determined such that the service distance is maximised, the service density is minimised, as defined in Section 4, and at the same time the following constraints are satisfied:

- The reliability constraint,
- The latency constraint,
- The cable length constraint (which depends on the cabling technology).

The latency and reliability constraints are the targets outlined in Table 2. The cable length constraint is based on signal degradation over the link between two CN nodes. For fibre links, the maximum distance between two CN nodes before signal degradation occurs depends on a number of factors including the properties of the cables and whether single mode or multi-mode operation is used [18]. For the scope of this analysis, multi-mode transmission is considered, which allows for a maximum distance of approximately 2 km. For single mode operation, the maximum distance can be much larger (e.g. 40 km to 100 km depending on the capacity of the link among other factors) [19]. However, single-mode operation is generally a more expensive choice than multi-mode operation. Clearly, this analysis is still adequate should single-mode operation was to be assumed, as long as the appropriate parameters are chosen.

As the maximum inter-node distance decreases, more forwarding nodes are required to reach the same service distance. However, increasing the number of forwarding nodes increases the latency due to additional processing at each forwarding node.

In order to increase the service distance and reduce the service density, the optimum number of forwarding nodes should be determined, where the optimum number of forwarding nodes is the maximum N_i that satisfies the above constraints.

The service distance is given by

$$d_{service} = r_c + d_{CN} \tag{11}$$

where r_c is the cell radius of the RAN node and d_{CN} is the distance from the RAN node to the AS. This is illustrated in Figure 10. If the RAN deployment consists of a distributed antenna system then the service distance also includes the latency over the fronthaul. For example, if CPRI is used over the fronthaul and the maximum distance between the antenna and the AP is 20 km then a maximum fronthaul latency of $100\ \mu s$ can be added to the RAN latency.

The problem can be formulated as follows:

$$Maximize\ d_{service} = r_c + d_{CN} \tag{12}$$

Subject to

$$t_{E2E} \leq t_{E2E}^{*}$$
$$R_{E2E} \geq R_{E2E}^{*}$$

$$L_{cable} \le L_{cable}^*$$

The parameters t_{E2E}^* and R_{E2E}^* are the target E2E latency and the target E2E reliability requirements, respectively. For the purpose of this analysis, the considered targets are listed in Table 2. The parameter L_{cable}^* is the maximum cable length.

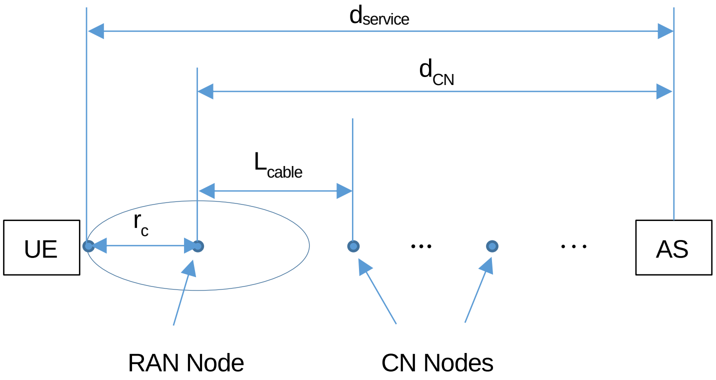


Figure 10: Service Distance = cell radius + core network distance.

7 NUMERICAL ANALYSIS

In this section, the latency analysis and reliability analysis described in Sections 6.2 and 6.1, respectively, are each performed separately in order to determine the factors and trade-offs affecting the E2E performance. The analysis is then followed by the joint optimization of latency and reliability requirements, outlined in Section 6.3.

As described in Section 6.3, the goal is to maximize the service distance, i.e., the distance between UE and AS, as defined in Equation 11, where r_c is the radius of the cell (the RAN is assumed to be centralised, i.e., not distributed), and d_{CN} is the distance from the AP to the AS.

7.1 Latency Analysis

The methodology for the latency analysis described in Section 6.2 is applied to the three target uses cases described in Table 2. Only the latency constraint is considered in this section. The values used for the parameters are numerical examples that reflect as much as possible realistic deployment scenarios. However, any other value can also be used, depending on what scenario needs to be represented.

For Target 1, the assumptions are as follows:

- Latency in the RAN, $t_{RAN} = 10$ ms
- Cell radius $r_c = 250$m
- Packet size, $P_{size} = 1500$ Bytes
- The capacity for each link = 1 Gbps
- Fraction of the link capacity allocated to the target use case $\alpha C_{link} = 700$ Mbps (i.e. $\alpha = 0.7$),
- Node processing delay, $t_{process} = 200$ μs [20] [2]

As the cell radius is fixed by the chosen RAN solution, the goal is to maximize the CN distance d_{CN} in Equation 11. The maximum CN distance d_{CN} that allows to satisfy the E2E latency requirements (given a fixed r_c) versus the number of forwarding nodes is illustrated in Figure 11. As can be seen, as the number of forwarding nodes increases, the CN distance clearly has to decrease. If the fraction of the link capacity reserved for the target use case is increased further (e.g. if $\alpha > 0.7$) there is some improvement. However, there will be less capacity available for other traffic.

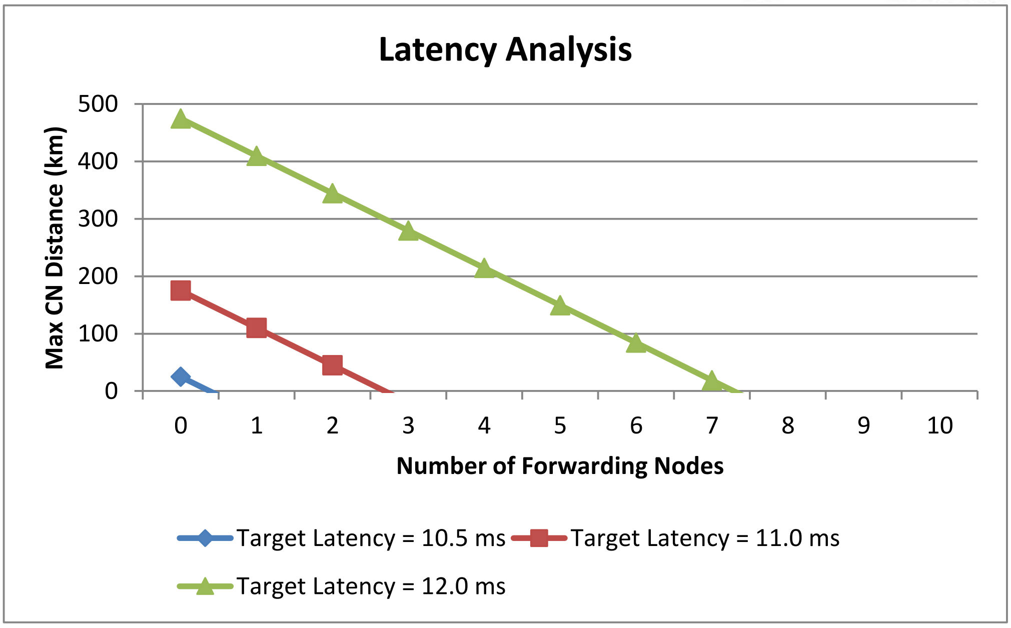


Figure 11: Latency analysis for Target 1.

For Target 2, the assumptions are as follows:

- Latency in the RAN, $t_{RAN} = 5$ ms
- Cell radius r_c = 250m
- Packet size, $P_{size} = 40$ Bytes
- The capacity of each link = 1 Gbps
- Fraction of the link capacity allocated to the target use case $\alpha C_{link} = 100$ Mbps (i.e. $\alpha = 0.1$),
- Node processing delay, $t_{process} = 200$ µs

The maximum distance in the CN versus the number of forwarding nodes for Target 2 is illustrated in Figure 12. In this case, the fraction of the link capacity reserved for the Target 2 (i.e. $\alpha = 0.1$) is much lower than for Target 1. Increasing the link capacity beyond 100 Mbps does not provide additional improvement, since the packet size is much smaller than for Target 1. In this case, more capacity is available for other traffic.

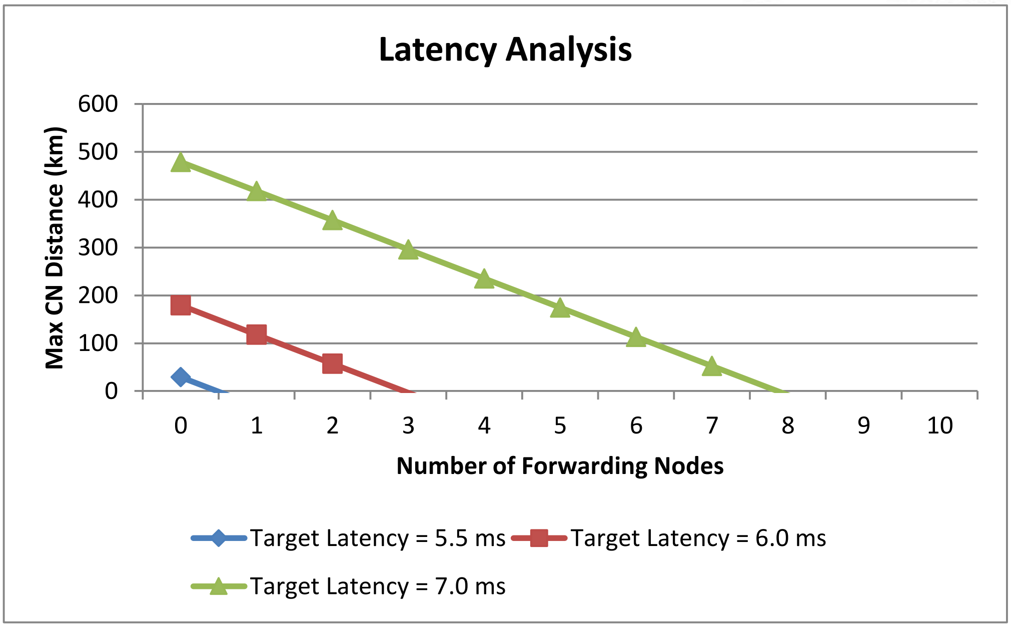


Figure 12: Latency analysis for Target 2.

For Target 3, the assumptions are as follows:

- Latency in the RAN, $t_{RAN} = 1$ ms
- Cell radius $r_c = 250$m
- Packet size, $P_{size} = 40$ Bytes
- The capacity of each link = 1 Gbps
- Fraction of the link capacity allocated to the target use case $\alpha C_{link} = 100$ Mbps (i.e. $\alpha = 0.1$),
- Node processing delay, $t_{process} = 200$ µs

The maximum CN distance versus the number of forwarding nodes for Target 3 is illustrated in Figure 13.

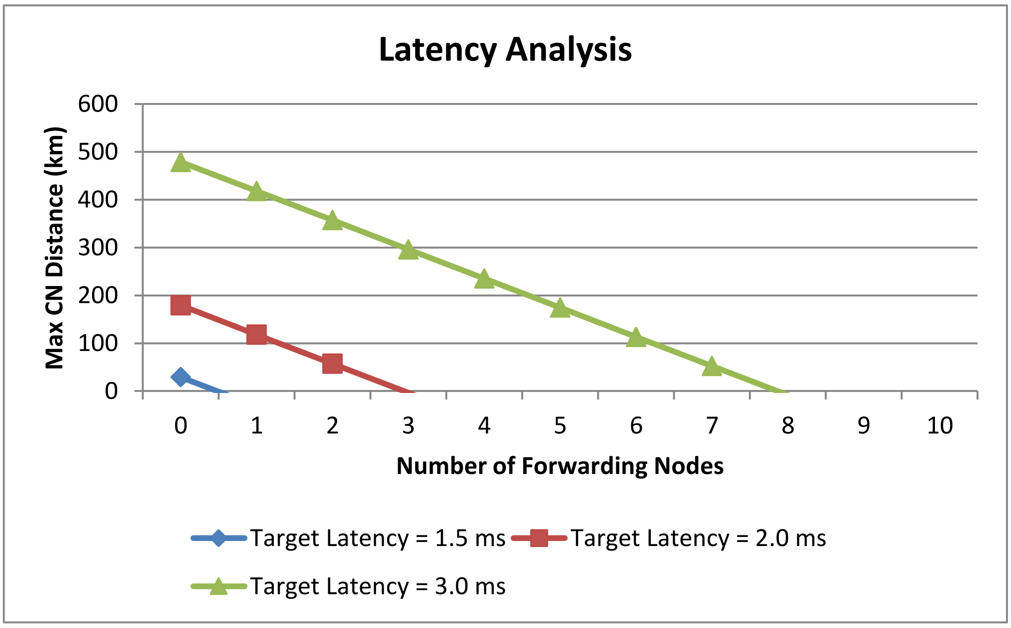


Figure 13: Latency analysis for Target 3.

Although the processing delay at each node is assumed to be fixed in the above examples, in reality, the delay depends on a number of factors. For example, if there is congestion in the network, there will be a queueing delay. If additional security is required in the CN then there will be a delay associated with encryption and decryption.

The E2E latency performance depends on both the 3GPP delays as well as the non-3GPP (N3GPP) delays. The trade-off between the delays in 3GPP and N3GPP is illustrated in Figure 14.

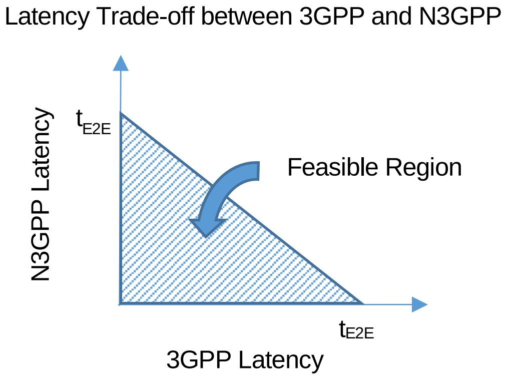


Figure 14: Latency Trade-off between 3GPP and N3GPP

The shaded area in the graph is the feasible region, where the target E2E latency is satisfied. The line on the right border of the region is where the actual E2E latency is equal to the target E2E latency.

The required latency target for the 3GPP portion of the latency budget depends on the capability of the UE and the AS. As the capability of the UE and the AS improve and the latency consumed by the UE and AS decrease, the requirements for the 3GPP portion can be relaxed.

The requirements for the N3GPP portion of the latency budget can be determined by calculating the total additional processing delay, which is given by

$$t_{add} = t_{UE} + t_{AS} \tag{13}$$

The additional processing delay at the UE, t_{UE} and at the AS, t_{AS} that can be accommodated for a given number of forwarding nodes can be calculated from Equations 7 and 8, given that the RAN latency is here considered as a fixed input.

The additional processing delays for Targets 1, 2, and 3 are illustrated in Figure 15, Figure 16 and Figure 17, respectively.

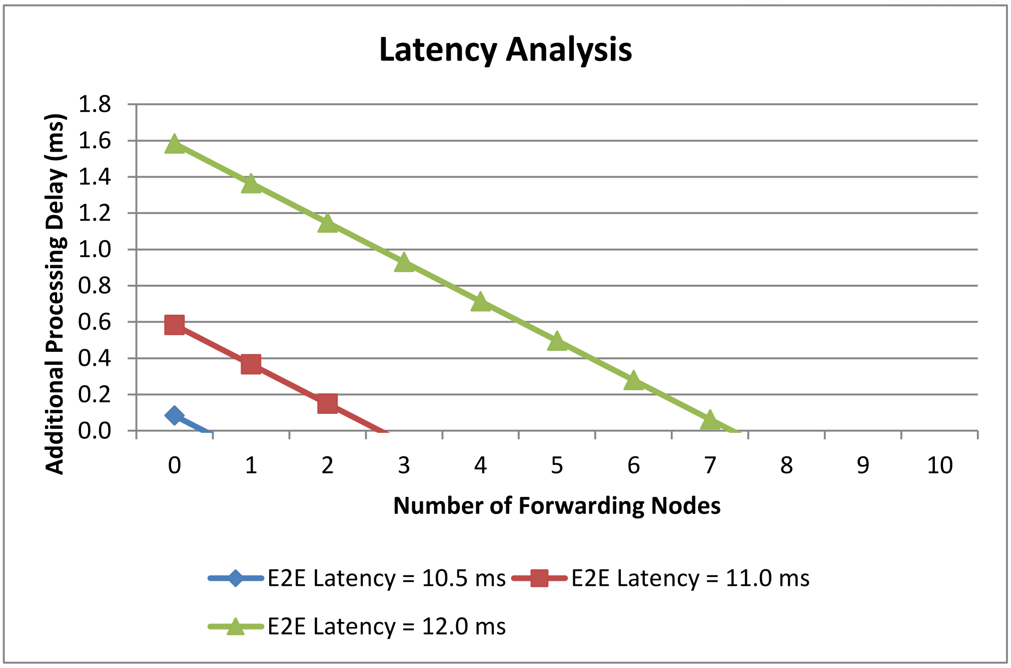


Figure 15: Additional Processing Delay for Target 1.

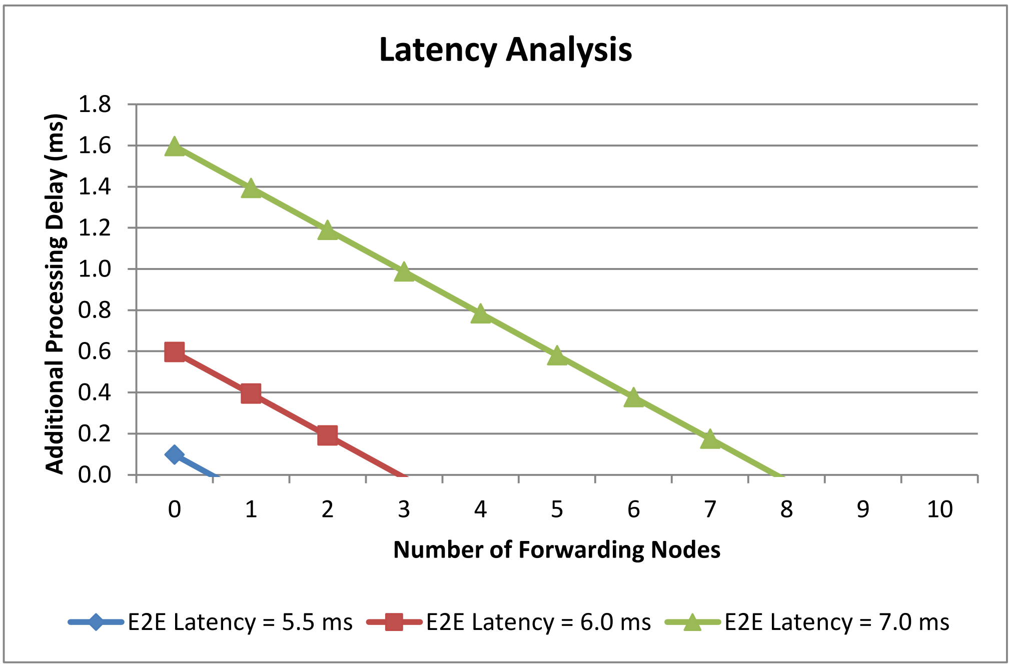


Figure 16: Additional Processing Delay for Target 2.

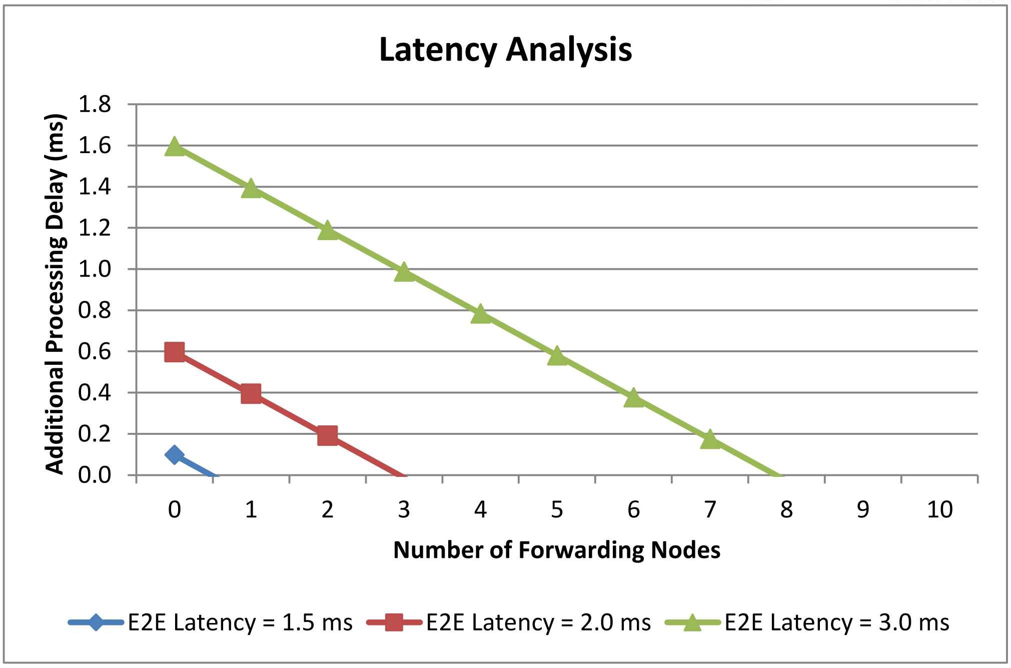


Figure 17: Additional Processing Delay for Target 3.

To determine whether the constraints on additional processing delays identified for the different targets are feasible, realistic processing delays for the N3GPP protocol layers at the UE and AS can be further analysed. The N3GPP delays are illustrated in Table 3. An E2E service session has multiple phases that should be differentiated:

- Session establishment
- Ongoing session
- Session tear down

Table 3 summarizes the delays that are associated with session establishment and on-going session.

Each of these phases have unique requirements, and different components that contribute to latency/reliability. This differentiated approach has been used already for voice services. The call setup time has a different requirement from tolerable transmission latencies during the session.

For the use cases listed in Table 1, the latency requirements refer to an on-going session only. For use cases like AR/VR, remote control, tactile interaction, etc. 100s of milliseconds or even seconds can be tolerated for session setup, whereas latencies around 10ms apply to the on-going session. For electricity distribution, these use cases are generally very long running sessions, which are configured once and then run continuously over multiple years. The session setup requirements can then be very relaxed (in the order of minutes). It is only during an on-going session that low latencies are required.

This differentiation of session setup and ongoing session stretches over 3GPP and non-3GPP domains. E.g., for session setup this could start all the way from registration on a network (e.g. when a new Smart Grid intelligent electronic device is installed at a substation and configured for power line protection).

Hence, for session setup, no quantitative analysis is needed as it is not a bottleneck when planning the edge node deployment.

For the N3GPP latencies during a session, the following components (a subset of components from Table 3) apply:

- Encryption / decryption at end-points or at intermediate security gateways outside the 3GPP domain.
- Congestion avoidance, as there is continuous probing and guesswork of the on-going session throughput.
- Encapsulation/decapsulation at end-points. Possibly encapsulation / decapsulation could happen at gateways that bridge multiple network domains beyond 3GPP. However, in the context of reliable and low latency services, no intermediate network hops are assumed to exist, and the AS is placed on the 3GPP edge.
- Packet forwarding, which, if the AS is on the 3GPP edge, should be small.
- Transport layer multiplexing, which happens at the end points.

An estimate of the actual N3GPP latency associated to an on-going session is illustrated in Table 4 [21] [22]. This latency represents processing delays associated with one end point (i.e., either UE or AS). Note: this does not include network round trips for handshakes and lookups [21].

From the results shown in Figure 15 Figure 17, if both the UE and AS incur the maximum $45\ \mu s$ latency then the total additional latency combining the processing at both nodes is $90\ \mu s$. This is within the maximum values that are acceptable for all targets, as illustrated in Figure 15 Figure 17.

Table 3: Components of Non-3GPP Processing Delays.

Phase	Component	Optional	Freq.	Example process**
Session establishment	Connection establishment	Yes	Once	Lookup IP + bind socket + lookup MAC + encapsulate + handshake
	Congestion avoidance	Yes,	Multiple	ACKs, SACKs, increase/decrease of segment transmission, retransmission
	Connection Securement	Yes	Once	TLS Authentication + crypto negotiation + Key Exchange
Ongoing session	Encryption/decryption	Yes	Per packet	Data encrypted , data decrypted
	Congestion avoidance	Yes,	Multiple	ACKs, SACKs, increase/decrease of segment transmission, retransmission
	Encapsulation/decapsulation	No	Per packet	Unpacking the payload of Lx from an Lx-1 packet
	Packet Forwarding	Yes	Per packet	Interim router allocates packet to flow based on headers/labels (MPLS, IP)
	Transport layer Multiplexing/ demultiplexing	Yes	Per stream	QUIC stream multiplexing/demultiplexing

* UDP is connectionless, but still requires a logical connection via socket binding

** items in red incur network or end-to-end metadata signalling, and hence propagation/transmission delay

Table 4: Non-3GPP Processing Delays for an end-point (UE or AS).

Minimum	Median	99.9th	Tail	Observation Period
3.9μs	4.5μs	21μs	45μs	1M packets

7.2 Reliability Analysis

Using the methodology described in Section 6.1, the maximum number of forwarding nodes can be determined based on the reliability requirement (i.e., without considering the latency requirement).

Table 5 illustrates the number of forwarding nodes that can be used to satisfy the E2E reliability requirement of 99.99% (i.e. 4 nines) assuming that the RAN reliability is 99.999% (5 nines) for a range of given link and node probabilities of failure. Table 5 is illustrated graphically in Figure 18.

The considered probabilities of link and node failure are provided by [16] [17].

Table 5: Maximum number of forwarding nodes that can be deployed in order to satisfy the E2E reliability target of 99.99%, given a RAN reliability of 99.999%, and given a pair of values for probability of link and node failure.

Probability of Link Failure	Probability of Node Failure		
	1.00E-07	1.00E-06	1.00E-05
1.00E-06	88	48	7
2.00E-06	46	31	6
3.00E-06	30	23	5
4.00E-06	23	18	5
5.00E-06	18	15	4
6.00E-06	15	13	4
7.00E-06	12	11	4
8.00E-06	11	9	3
9.00E-06	9	8	3
1.00E-05	8	7	3

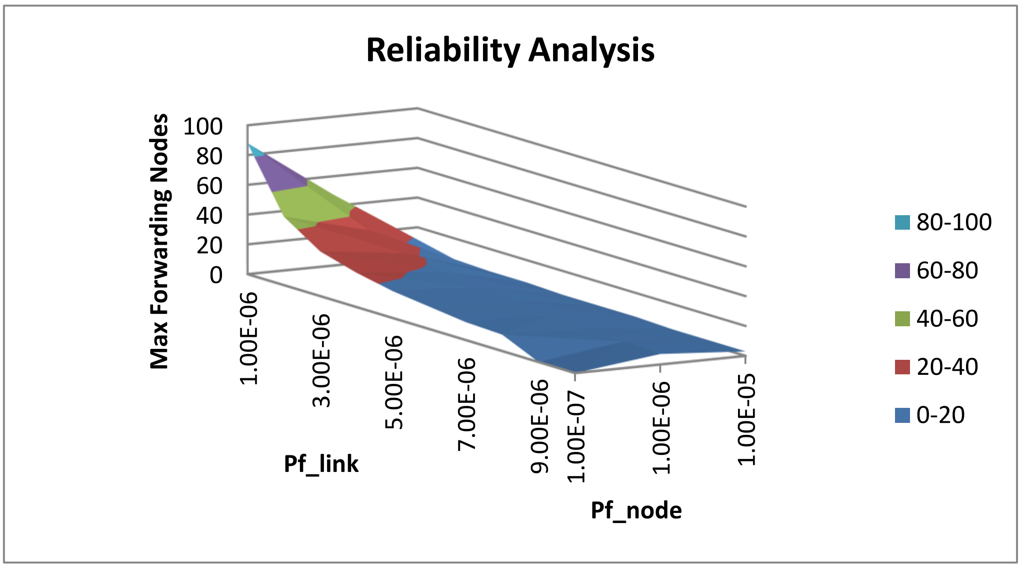


Figure 18: Maximum number of forwarding nodes that can be deployed in order to satisfy the E2E reliability target, given a RAN reliability of 99.999%.

There is a trade-off between the reliability in the RAN and the CN. By increasing the RAN reliability, the reliability in the CN can be decreased. The trade-off is illustrated in Figure 19.

In the analysis, CN and RAN are two components in the E2E reliability assessment and, as they are in series, the product of the individual reliabilities determines the E2E reliability. This is shown in Figure 19 for failure probabilities of 1e-3, 1e-4 and 1e-5.

From Figure 19, the following conclusions can be made:

- Each reliability for RAN and CN must itself be larger than the E2E reliability target.
- If one of the reliabilities (CN or RAN) is more than ~1 order of magnitude more reliable than the target E2E reliability, then its contribution becomes almost negligible. The required reliability of the other component needs to be only slightly more than the E2E reliability target (i.e. it becomes a straight line in the figure).
- Only if the reliability of both of the CN and RAN components is within up to approximately 1 order of magnitude from the target E2E reliability, a more careful assessment of the reliability components is needed to reach the target E2E reliability (i.e. the curved part in the figure).

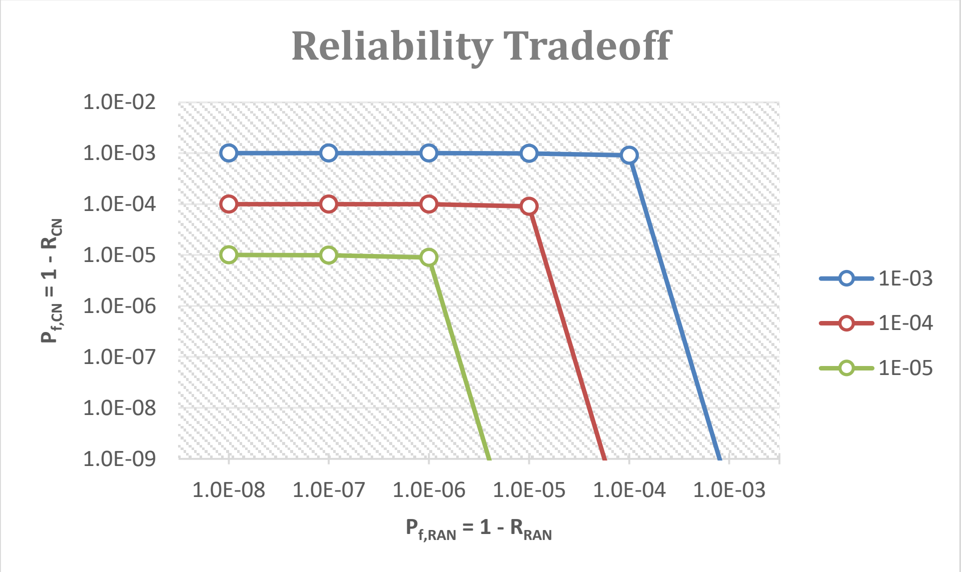


Figure 19: Reliability Trade-off.

7.2.1 Redundancy

Redundancy can be used to improve the reliability in the RAN and the CN without impacting the latency. The performance results for packet duplication in the RAN are illustrated in [23].

The impact of redundancy on the reliability is illustrated in Table 6. For example, if the reliability in the RAN using a single link is 99.999% (5 nines) then a reliability of 99.9999% (6 nines) can be achieved by using two redundant paths/links (i.e. packet duplication). In this case, the redundant paths provide reliability that significantly exceed even the target (i.e. 1e-10 for a requirement of 1e-6). Hence, the reliability requirement on the two redundant paths can be relaxed to achieve the overall 5-nines reliability.

Similarly, the reliability in the CN can also be increased using redundant paths. If the target CN reliability is 6 nines and the single path reliability is between 3 nines and 5 nines then two redundant paths are required to satisfy the target reliability.

Although redundancy requires resources on multiple links, the performance results in [23] illustrate that, under certain conditions such as during handover, the resource efficiency is improved when packet duplication is activated.

Table 6: Impact of Redundancy on Reliability.

Single Path Reliability	Single Path Reliability (#nines)	Path Redundancy	Overall Reliability	Overall Reliability (#nines)
90.000%	1	6	99.999900%	6
99.000%	2	3	99.999900%	6
99.900%	3	2	99.999900%	6
99.990%	4	2	99.999999%	>6
99.999%	5	2	100.000000%	>6

Redundancy can compensate for node failures as well as link failures. The UE and the AS can be considered as nodes in the E2E path. Both the UE and AS can be a single point of failure. Fault management techniques may be required at all nodes in order to satisfy the E2E reliability requirements. If the reliability of the UE and AS are considered, it may be necessary to duplicate the UE and AS to ensure that there is no single point of failure. In this case, multiple instances of the UE and AS functions can be instantiated on separate nodes. The multiple instances should be synchronized. Stateless functions can be used, where the state information is stored externally. This reduces the synchronization effort to only the storage of the state information.

7.2.2 Dual Connectivity and Carrier Aggregation

Carrier Aggregation (CA) was introduced in 3GPP to allow a UE to simultaneously transmit and receive data on multiple component carriers from a single eNB. The main advantage of CA is the increase in user throughput as the aggregate bandwidth is increased.

Dual Connectivity (DC) was introduced in 3GPP to allow a UE to simultaneously transmit and receive data on multiple component carriers from two cell groups via master eNB (MeNB), or master node (MN), and secondary eNB (SeNB), or secondary node (SN). The advantage of DC is that it can increase user throughput, provide mobility robustness and support load-balancing among eNBs.

In 3GPP NR, DL and UL packet duplication is supported as a tool to enable redundancy and hence to provide high reliability with ultra-low latency. Packet duplication can be performed in either Carrier Aggregation (CA) or in Dual Connectivity (DC). In both CA and DC, there is only one Packet Data Convergence Protocol (PDCP) layer in the UE and one in the RAN. The packets are duplicated in the transmitting PDCP entity. Duplicate packets are removed in the receiving PDCP entity. Some architectural enhancements and performance analysis using packet duplication for URLLC are provided in [24].

The RAN solution for providing redundancy is illustrated in Figure 20, where two redundant paths are created between the UE and a MN and a SN.

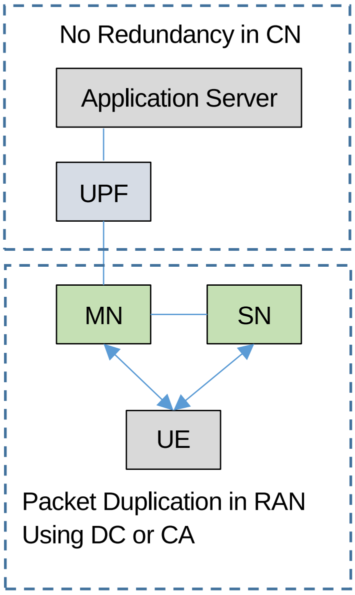


Figure 20: Packet Duplication in RAN using Dual Connectivity or Carrier Aggregation.

In this case, there is only one connection to the CN through the MN. This solution may be sufficient in some cases, such as when the AS is collocated with the RAN node. However, if the AS is located further in the core network then the reliability of the CN will impact the E2E reliability. If a single path in the CN cannot meet the required reliability target then redundant paths are necessary. An enhanced tunneling protocol may be required between the RAN and the CN that allows for redundant parallel transmissions on different paths. Some enhancements for the CN have been identified in [25] and will be addressed in [26].

In order to provide redundancy in the CN and to further improve the reliability, there are several RAN architecture options for dual connectivity (DC) to consider.

In the first option, which is illustrated in Figure 21Figure 20, there are two connections to the CN through one node. The UE is connected to a MN and a SN. However, only the MN is connected to the CN. Reliability in the RAN is achieved by packet duplication. A split bearer can be configured with duplication for the Protocol Data Unit (PDU) session handling URLLC traffic.

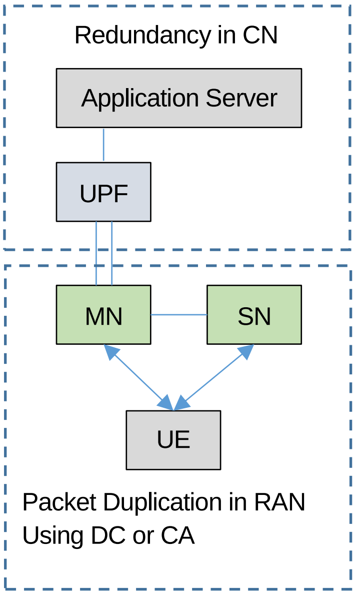
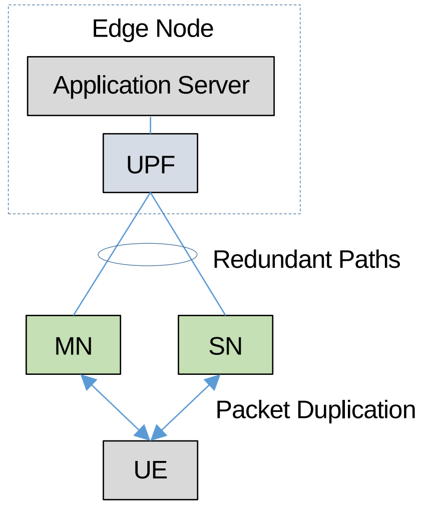


Figure 21: Redundancy in the RAN and the CN. There are two connections to the CN through one RAN node (e.g. MN).

In order to improve the reliability in the CN for this option, enhancements are required to allow redundant paths between the RAN node and the CN. Packets would have to be duplicated and removed in the MN and the UPF. The disadvantage of this approach is that the MN is a single point of failure in addition to the UE and the AS. Also, the latency over the interface between the MN and SN (i.e. X2 or Xn) may impact the E2E latency.

In the second option, illustrated in Figure 22, there are two separate connections to the CN (one from each access point). There are two redundant paths from the UE to the AS. Only the UE and the AS are single point of failures in this case. Two independent traffic flows are transferred between the CN and the UE via the RAN. One flow is mapped to a Master Cell Group Bearer terminated in the MN; the other flow is mapped to a Secondary Cell Group Bearer terminated in the SN. This approach addresses the issue with the additional latency over the interface between the access points.



DC Architecture with Two Connections to the CN

Figure 22: Redundancy Using DC with Two Connections to the CN.

However, some enhancements are required. Data duplication for this mode would need to happen outside the Access Stratum (see Figure 22). A UPF would need to enable the duplication mapped to different bearers. Similarly, a UE would need to duplicate the packets at a layer above the radio protocols and map them to different bearers. A function on the receiving side is also needed that filters out duplicate messages, if it should not be exposed to the application. This duplication and duplicate deletion are not currently standardized in 3GPP.

The setup in Figure 22 can be realized by establishing one PDU session with two N3 tunnels, which can already be achieved in 3GPP when the UE is configured for DC. Some changes are required to add the packet duplication and removal function in the UPF and the UE. A new QoS flow mapping rule should also be specified for mapping the duplicate flows to different tunnel end points (e.g. original flow to MN and duplicate flow to SN).

In order to improve resource efficiency in the RAN, dynamic control of packet duplication should be supported in the RAN. If packet duplication is activated then the two N3 tunnels are configured from the UPF to two separate nodes (i.e. MN and SN). Otherwise, if packet duplication is deactivated in the RAN, but still required in the CN, then one of the N3 tunnels can be reconfigured so that there are two N3 tunnels to the same node (i.e. the node with the best link) as in Figure 21. Also, if packet duplication is required in the RAN and not required in the CN then the solution in Figure 20 can be used.

Alternatively, the packet duplication and removal function could be outside the 3GPP domain. For example, it can be provided by a transport protocol. In case of Ethernet traffic, TSN defines such a Frame Replication and Elimination for Reliability (FRER) function in 802.1CB. In case of Ethernet, the duplication could happen at the AS outside the 3GPP domain, and the corresponding FRER function could be at the UE (but outside the 3GPP domain). The same duplication schemes can be done based on IP protocols. The IETF working group on Deterministic Networking has in its charter the goal to define such an IP based protocol (see https://tools.ietf.org/html/draft-ietf-detnet-problem-statement-03).

This option would have small impact on 3GPP. Since, two separate PDU Sessions are established for a UE to the same UPF, the PDU session establishment procedure should be modified so that the same UPF is selected for both PDU sessions. Data would be duplicated by a higher layer function and would be transmitted via the two PDU Sessions. The two PDU Sessions would need to be established independently, through the Master Cell Group Bearer and Secondary Cell Group Bearer to ensure that one connects via the MN and the other one via the SN in the dual connectivity setup. No user plane tunnel protocol between UPF and UE needs to be specified in 3GPP.

However, in this case, if packet duplication is deactivated in the RAN then it is also deactivated in the CN. This solution is not as efficient in terms of resource usage as in the previous solution, since redundancy in the CN can only be achieved if the packets are also duplicated in the RAN. If the UE is not configured with DC then the E2E reliability may not be satisfied if redundancy is required in the CN.

The case where the UPF is collocated with the AS and the MN and SN both have a connection to the same UPF is illustrated in Figure 22. If the UPF is not collocated with the AS then it may be desirable to have separate UPFs for each path. The two architecture options for this case are illustrated in Figure 23.

The architecture on the left is the case where there are two connections to CN from one node (e.g. MN) to two different UPFs. The packet duplication and removal function is performed by the PDCP entity in the UE and the RAN node hosting the PDCP. If redundancy is also required in the CN then the MN and the AS (or a UPF collocated with the AS) are required to perform the packet duplication and removal function.

The architecture on the right is the case where both the MN and SN have a connection to the CN through different UPFs. In this case, the packets that are duplicated in the RAN are not removed in the RAN. The duplicate packets can be removed in the application in the AS (or UPF collocated with the UPF) and the UE.

In order to improve resource and energy efficiency, dynamic control of packet duplication in the RAN is included in NR. Packet duplication can be activated only when it is required (i.e. the channel condition for the best link is below a threshold). Although the redundancy in the CN can be controlled semi-statically rather than dynamically, the dynamic control of packet duplication in the RAN may impact the redundancy in the CN in some scenarios.

Since the requirements for redundancy in the RAN and the CN depend on the respective target reliabilities, the redundancy decision should be performed independently by the RAN and the CN. However, the solution on how to achieve the redundancy should be selected based on both redundancy requirements.

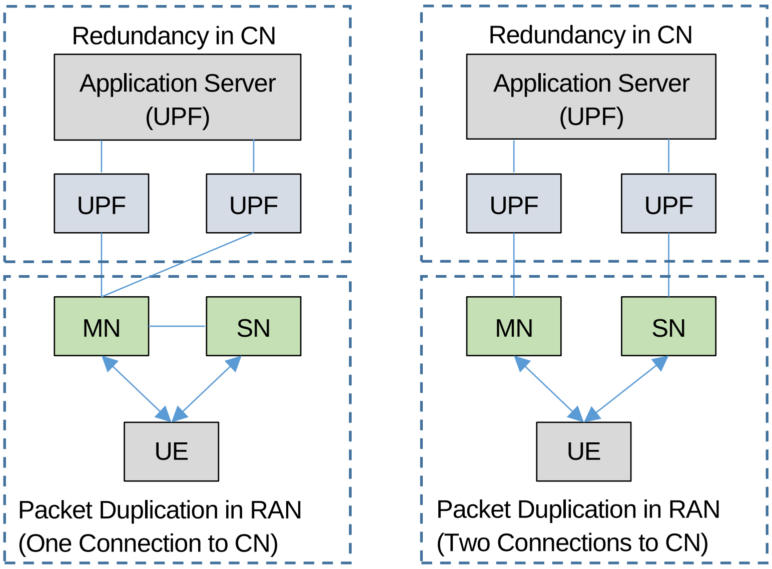


Figure 23: Redundancy in the RAN and CN with two UPFs.

There are two ways to achieve the solutions in Figure 23. In the first solution, there is one PDU session with multiple N3 tunnels. In the other solution, there are two separate PDU sessions.

One PDU Session:
For the entire setup in Figure 23, one PDU session can be established with two N3 tunnels.

The setup on the left is achieved by configuring both N3 tunnels to connect to the same RAN node through different UPFs. This configuration can be used when packet duplication is required in the CN and not required in the RAN. If packet duplication is required in the RAN, it can be activated using the existing solution for the packet duplication in the RAN can be used. In this solution, packet duplication in the CN is performed at the MN and at the UPF that is collocated with the AS. Packet duplication is performed independently in the RAN and the CN and can be activated only when it is required.

If packet duplication is required in both the RAN and the CN then one of the N3 tunnels can be configured to the second RAN node as in the case on the right side of Figure 23. In this case, the packet duplication is performed by the UE (above the AS layer) and the UPF collocated with the AS. Alternatively, the packet duplication and removal can be performed at the UE and AS (outside of the 3GPP domain).

Configuring two N3 tunnels within one PDU session and one UPF can already be performed in the standard today. This is used to support the DC architecture in the RAN. The only modification that is required is to enhance the PDU session establishment procedure to select two different UPFs. The two UPFs can be configured with N3 tunnels to the same or different RAN nodes. Duplicate flows should be mapped to the different N3 tunnels.

The entire solution can be achieved by configuring the multiple N3 tunnels to both RAN nodes and configuring rules for mapping the duplicate QoS flows to the different tunnels. The rules can take into account the requirements for redundancy in both the RAN and the CN.

Two PDU Sessions:
In another alternative, the entire solution in Figure 23**Fehler! Verweisquelle konnte nicht gefunden werden.** can be supported by establishing two separate PDU sessions. The solution on the right hand side of Figure 23 consists of two independent paths between the UE and the AS through two different APs and two different UPFs. This solution can be achieved by establishing two separate PDU sessions between the AS and the UE; one PDU session is mapped to a MCG Bearer, the other one to a SCG Bearer. For the solution on the left hand side, the two PDU sessions are mapped to the same RAN node.

In this case, the packet duplication is performed outside of the 3GPP domain, (e.g. in the application layer). A copy of every packet is delivered via one or both APs and into each PDU session.

However, in this alternative, if packet duplication is deactivated in the RAN then it will be deactivated in the CN. This solution cannot support dynamic control of packet duplication in the RAN, which is not an efficient use of the RAN resources. For example, if the UE is no longer configured for DC or packet duplication is deactivated and duplication is required in the CN then the UE must send duplicate packets on two different PDU sessions to the same RAN node in order to satisfy the redundancy requirement in the CN.

In contrast to Figure 21, the packet duplication and removal function can be performed at the application layer or at the transport layer. The AS duplicates messages and sends them via the different PDU sessions. Similarly, the application in the UE duplicates messages and sends them to different PDU sessions. The receiving sides in the application in the UE and the AS would then remove duplicates.

Although it is possible to establish two separate PDU sessions in the standard today, there is no guarantee that two different UPFs are selected when two separate PDU sessions are established. In order to ensure there are two distinct UPFs and paths between the UE and AS, some modification to the PDU session establishment procedure are required. For example, one PDU session can be established with two different UPFs and two separate N3 tunnels.

In this solution with two PDU sessions, if packet duplication is deactivated in the RAN, but is still required in the CN then one of the N3 tunnel endpoints should be modified so that both tunnels terminate at the same RAN node. This solution then becomes the same as the previous case where the MN and AS (i.e. the UPF collocated with the AS) perform the packet duplication and removal function.

7.2.3 Control Plane Reliability

In Section 6.1, Equation 6 has been introduced to calculate the joint reliability of two redundant paths through the RAN.

This formula is based on the assumption that the RAN paths are independent. However, even with dual connectivity and usage of two disjoint (user-plane) paths, the reliability of the paths remain correlated. There is a common point of failure for both paths, which is the common control plane.

The UE has a single control plane Radio Resource Control (RRC) connection to the MN. Also, the radio link failure procedure is bound to the primary link. If user-plane data is sent via two redundant paths through the MN and the SN, when the radio link of the primary link deteriorates it will trigger a radio link failure (RLF) even if there is an error free redundant link. As a result of RLF, the connection of the UE (towards both MgNB and SgNB) is terminated and re-established.

This has been identified in [27] and will be addressed in 3GPP Rel-16 (see a WI proposal RP-180456).

It should be noted that also in the CN there is some dependence on a control plane. For example, if the Home Subscriber Server (HSS) or the Access Management Function (AMF) fail, the redundant paths may not be able to be maintained (e.g. in case of mobility). However, HSS or AMF failure may not lead to an immediate interruption and thereby may allow for node restoration during the grace period before the path failure.

7.3 Joint Analysis of Latency and Reliability

Based on the methodology described in Section 6.3 on the joint reliability and latency analysis, a numerical example that provides deployment recommendations for the RAN and CN can be determined, based on a given set of assumptions.

It is important to emphasize that the outcome of this exercise is dependent on the goal, and on the chosen inputs. The goal, as stated in Section 6.3, is to maximise the service distance, i.e., the physical distance between UE and AS in order to reduce the number of edge nodes required for a given application, or to provide a regional boundary within which a single edge node can deliver the service.

The inputs are a set of assumptions that have been made on RAN latency and reliability [2], node processing delays, link capacity, packet size, node and link failure probabilities, and cable length constraint, as defined in Section 6.3. All inputs are outlined in Sections 7.1 and 7.2. A different set of inputs will clearly lead to a different outcome, and a different optimisation goal will lead to a different methodology.

The numerical exercise is summarised in Table 7. The targets from Table 2 are used as inputs, and are also included in the first three columns of Table 7.

For each E2E latency and reliability target, the following outputs are given:

- **Maximum number of forwarding nodes**, i.e., hops, that can exist between the AP and the AS. As explained in Section 6.3, the forwarding nodes allow to extend the distance between AP and AS because they re-generate the signal travelling through fibre, but at the same time they add processing delay to the E2E communication chain. Hence, from the service distance perspective, it is advisable to introduce as many forwarding nodes as possible, whereas from an E2E latency perspective, it is advisable to introduce the least possible number of hops. The optimal number is thus the maximum possible value that allows to comply with the E2E latency requirement.
- The **maximum node and link failure rates** that can be tolerated in order to deliver the target E2E reliability, given the number of forwarding nodes calculated in the previous column.
- The **maximum physical distance** between AP and AS, given the maximum number of forwarding nodes.
- The **required redundancy** both in the RAN and in the CN, given the reliability of a single path. If the target RAN reliability cannot be satisfied with a single link, then packet duplication can be used to achieve the target reliability. Similarly, in the CN, if the target CN reliability cannot be achieved with a single path then redundant paths can be used.
- The number of APs that are served by a single AS.
- The **service distance**, considering the assumption made on cell radius (250m).

Table 7: Numerical analysis.

Target	Target E2E Reliability	Target E2E Latency (ms)	Max Fwd. Nodes	Maximum Node/Link Failure Rate	Max CN distance (km)	Required Redundancy		Service Density	Service Distance (km)
						RAN	CN		
1	99.9%	10.5	0	1e-5/9e-4	2	Target R_RAN = 99.99% 2 (if single path is 3 nines) 1 (if single path is 4 nines)	1 (if single path > 99.91)	64	2.25
		11	2	1e-5/3e-4 1e-6/3e-4	6			576	6.25
		12	7	1e-5/1.1e-4 1e-6/1.2e-4	16			4096	16.25
2	99.99%	5.5	0	1e-5/7e-5, 1e-6/9e-5	2	Target R_RAN = 99.999% 2 (if single path is 4 nines) 1 (if single path is 5 nines)	2 (if single path is 3 nines) 1 (if single path > 99.991)	64	2.25
		6	2	1e-6/1e5e-5	6			576	6.25
		7	7	1e-7/1.2e-5, 1e-6/1.2e-5, 1e-5/1e-6	16			4096	16.25
3	99.999%	1.5	0	1e-6/7e-6	2	Target R_RAN = 99.9999% 2 (if single path is 5 nines) 1 (if single path is 6 nines)	2 (if signle path is 4 nines) 1 (if single path > 99.9991)	64	2.25
		2	2	1e-6/1e-6	6			576	6.25
		3	7	1e-7/1e-7	16			4096	16.25

8 TUNNELING AND TRANSPORT LAYER PROTOCOLS FOR EXTREME REQUIREMENTS

The inter-dependence and inherent trade-off between latency and reliability impose strict constraints on any potential tunnelling protocols and transport layer protocols that can be used to support the target use cases with extreme requirements. The following describes the applicable scenarios and potential enhancements that may be required for the corresponding tunnelling protocols (within 3GPP domain) and transport layer protocols (outside 3GPP domain).

8.1 Tunnelling Protocol Enhancements for Extreme Requirements

Tunnelling protocols are generally used to augment packet forwarding capabilities on a point-to-point basis without modifying the under-laying lower layer protocols. In RAN and CN, the tunnelling protocols are utilized in two scenarios namely, i) between RAN and UPF and ii) between UE and UPF.

For satisfying extreme requirements within the 5G Core Network (5GC), the existing user plane GPRS tunnelling protocol (GTP-U), used over the N3 (RAN-to-UPF) and N9 (UPF-to-UPF) interfaces, can be applied in some scenarios without any changes. For example, if the packet duplication and removal function is performed outside of the 3GPP domain then the existing GTP-U protocol can be used.

If the packet duplication and removal function is performed within the 3GPP domain at the UE and the UPF collocated with the AS then some minor modification are necessary. It may be necessary to enhance the tunnelling protocol with in-built redundancy capability in order to satisfy higher reliability requirements. In this case, the enhanced tunnelling protocol in the CN should handle the transportation of duplicate packet flows over independent tunnels with distinct tunnel endpoint identifiers (TEIDs) per transport bearer. At the receiving entity in RAN (in DL) or UPF (in UL), the duplicates can be detected based on common sequence numbers and removed by the tunnelling protocol prior to sending the packets to the higher layers.

In order to satisfy the extreme requirements above the access stratum (AS) layer but within 3GPP, packet duplication and proactive transmission (without ACK/NACK feedback) can be handled by an E2E tunnelling protocol, established between the UE and UPF at the non-access stratum (NAS) layer. Conventionally, the NAS layer is used to support QoS flows associated with PDU sessions based on filtering rules configured by the control plane (CP) functions in CN. Applying similar principles in the E2E tunnelling protocol in UL, the IP packets from upper layers are duplicated and marked with different QoS flow identifiers in the UE prior to assigning them to different DRBs in the AS layer in the RAN. In the CN, the QoS flow markings on the packet headers can be used to assign the duplicate packets to different GTP-U tunnels, which are subsequently detected and removed at the NAS layer in UPF. Similarly, in the DL, the E2E tunnelling protocol in the NAS layer can perform packet duplication/marking in UPF and packet detection/filtering at the UE.

To ensure higher resource efficiency, it should be possible for any enhancements made to both CN and E2E tunnelling protocols to support automatic fall-back mechanisms that results in varying levels of redundancy without having to incur delays due to tunnel setup and modifications. The redundancy levels, defined in terms of the number of independent tunnels, can be made adaptable while ensuring that any path modification results in a graceful change to the end-to-end reliability and latency. For example, in the event where any one of the tunnels were to experience improvement in reliability or latency, duplicated transmission on other tunnels may be disabled to conserve resource usage and minimize loading without interrupting the ongoing packet forwarding. Likewise, new tunnels may be proactively added when the end-to-end reliability performance degrades below a certain threshold.

For the CN tunnelling protocol, the redundant tunnel adaptation capabilities can be supported by enhancing the existing tunnel management functionalities, handled by the GTP control protocol (GTP-C). Similar capabilities can be supported on the E2E tunnelling protocol using CP signalling at the NAS layer to activate and deactivate the redundancy levels based on performance monitoring and other triggers. For example, to satisfy E2E extreme requirements in UE mobility scenarios, CP signalling can be used to activate packet duplication and packet forwarding over 2 tunnels prior to handover and disabled after handover is completed.

8.2 Transport Layer Protocol Enhancements for Extreme Requirements

In the application layer, while the commonly used connection-oriented transport layer protocols such as TCP ensure end-to-end reliability, the high latency due to the initial handshaking and recovery from congestion conditions severely limits its applicability to only applications with best effort and delay-tolerant traffic characteristics. The stateful mechanism used in TCP based on in-sequence delivery and retransmissions results in increased complexity and higher transmission latency exceeding that of the maximum tolerable bound.

Alternatively, connectionless transport layer protocols such as UDP eliminate the need for handshaking and retransmissions. However, these improvements in latency come at the expense of reliability performance because UDP, while suitable for critical time-sensitive traffic, does not provide the means to recover from packet losses. In this regard, both of these conventional transport layer protocols are not suitable for the use cases under consideration where error-free reception of packets within the latency bound is of paramount importance.

For satisfying extreme requirements, the design considerations for a new transport layer protocol should jointly account for both latency and reliability while effectively balancing the trade-off between these two performance metrics. To this end, the new protocol should have multi-homing capability and inherently support transmission over multiple redundant paths. In this respect, the new protocol is similar to the existing multipath protocols such as MP-TCP and SCTP.

However, contrary to the existing protocols, the redundant paths should be used for transmission of duplicate packets rather than for achieving higher throughput and load balancing. Moreover, the selection of the redundant transmission paths should be performed such that the traffic flow on each path is isolated and uncorrelated on an end-to-end basis for ensuring maximum reliability.

The new transport protocol should also support stateless operation and allow proactive autonomous retransmission of packets on different paths without requiring explicit acknowledgement messages from the receiving entity. This approach can be applied to eliminate any additional RTT latency while at the same time, satisfy the end-to-end reliability target. Autonomous retransmission is similar to packet duplication. The decision to activate and deactivate the duplication is based on the reception of an ACK within a predefined duration on a single path. If the ACK was not received within the time window, the protocol can automatically start sending the retransmissions on a separate path. A potential transport layer protocol that may be applicable for extreme requirements use cases is the UDP-based multipath QUIC [28], however, further enhancements may be necessary to support simultaneous transmission of duplicate packets.

Note that even in the case when multiple UDP logical connections are established between the application layers in the UE and application server, there is no guarantee that the duplicate packets are not multiplexed in common tunnels and traverse via independent paths in the CN, thereby avoiding any single points of failures. In this regard, awareness of the UPF and tunnelling protocol configurations at the application and likewise, the awareness of the transport layer protocol configurations in the CN/NAS layer may be necessary for supporting the extreme requirements on an end-to-end basis.

In general, routing the packets to and from the application server via the Internet may lead to unpredictable latency and packet losses. Furthermore, using a centralized application server and locating it remotely in order to achieve wider service coverage can result in overloading conditions as well as packet queuing and processing related latency, both at forwarding nodes and the application server.

For the considered target use cases, a key aspect to take into account is the potential deployment of the application server in an edge node, located in close proximity to the RAN and CN. In another scenario, the application server may be hosted in a private network, directly accessible from the CN. In both of these scenarios, the generated packets need not traverse via the Internet and therefore, the flow control and congestion control techniques conventionally handled by the transport layer, can be moved to the application layer. This significantly simplifies the design of both the network and transport layer protocols, which in turn can be focused more towards achieving deterministic latency and consistent end-to-end reliability performance.

As a matter of fact, 3GPP provides support for both IP and non-IP based (e.g. Ethernet, unstructured) protocols at L3 in the 5GC. These enhancements, along with potential support for edge computing such as Multi-access Edge Computing (MEC), should be taken into consideration when optimizing the new transport layer protocol for satisfying the extreme requirements.

9 SATISFYING EXTREME REQUIREMENTS IN MOBILITY SCENARIO

Conventional handover (HO) in UE mobility scenario involves lengthy procedures and interactions with multiple functional entities in both the RAN and CN. The connection between the UE and the network is also interrupted during HO, where the link between the UE and source AP is terminated before a new link is established with the target AP, hence disrupting the ability to continuously transmit and receive data at all times. For uses cases such as AR/VR and mobile health where the extreme requirements need to be satisfied even when the UE is mobile, further enhancements are necessary in RAN and CN to not only eliminate HO related interruption but also to improve the transmission robustness.

A practical baseline technique to enable seamless and lossless HO in RAN would be the Make-Before-Break (MBB) procedure. Different from conventional HO procedure which relies on a single connection from the UE to the network at all times, MBB allows the connection to the source AP to be maintained even after the UE receives the HO Command to establish a connection with the target AP. Although significant reduction in the interruption time can be realized, for operation in NR RAN in 5G the current MBB technique is still not adequate to meet the low latency requirement due to the inability for the UE to simultaneously communicate with both source and target APs.

To satisfy the extreme requirements, it is vital to provide resilience to user plane (UP) packets during mobility by ensuring that there are no packet losses during HO. Here, enabling techniques such as packet duplication for the UP packets alone does not prevent failures nor would it directly improve the mobility robustness. However, performing packet duplication for both UP and control plane (CP) packets, as supported in NR RAN, can boost the overall reliability during mobility once the occurrence of any failure conditions is resolved. In this case, when either one of the links to source AP or target AP experiences link failure during HO, both the RRC and data connections can still be maintained.

The mobility management related procedures in the 5G Core Network (5GC) should also be aligned with the enhancements in the NR RAN. The exiting path change procedure in the CN where the CP function is required to modify the UP path from that of the source AP to target AP may no longer be adequate in light of the extreme requirements. In this regard, to ensure that the ongoing session in the CN is not interrupted during user mobility, three different session and service continuity (SSC) modes have been incorporated by 3GPP in the 5GC, each involving different session anchor node modification and traffic routing capability. Particularly, SSC Mode 3 enables MBB procedure to be supported in the 5GC where the connection to a new PDU Session Anchor can be established before terminating the connection to the existing PDU Session Anchor. For ensuring higher reliability, it may be necessary to enhance the existing session continuity procedure to support more robust transport layer protocols, redundant path and redundant UPF (re)selection techniques.

Additional changes for supporting extreme requirements during mobility may require support from application functions (AFs), located both within and outside the domain of network operators. While this may require interworking between the 3GPP and non-3GPP domains, the awareness of the capabilities external to 5GC such as AF mobility and multi-access edge computing can be useful for further optimizing the service performance on an E2E basis.

10 CONCLUSION

The aim of this work has been to identify factors that affect latency and reliability from an E2E perspective, i.e., considering both the 3GPP and non-3GPP domains, and to provide a methodology to optimise E2E network design in order to minimise cost of deployment whilst delivering the expected E2E performance. In this context, latency and reliability are considered jointly in establishing the design methodology, and the goal is to maximise the application server's distance from the user equipment to minimise the density of edge nodes or provide flexibility of deployment within a feasible area around the UE.

In order to test the proposed methodology, a numerical analysis is provided, which takes as input the radio access performance assessed in Deliverable 2.1 of this work [2], a set of E2E requirements, and a set of assumptions on node and link failure rates, and on processing and transmission delays. If the inputs changed, the methodology would still be valid.

From the results obtained from the analysis, the following remarks are in order:

- Extreme latency requirements can only be met where the AS is hosted in the operator network.
- The E2E reliability depends on the reliability of the UE, RAN, CN and AS.
- For a range of operating points, there is a trade-off between the reliability in the RAN and the reliability in the CN. If this range is significant, the following remarks apply:
  - The target reliability of the RAN can be reduced by improving the reliability in the CN. Conversely, the target reliability of the CN can be reduced by improving the reliability in the RAN. Each of them must be at least as high as the target E2E reliability.
  - The selected target reliabilities will impact the required network deployment.
- Redundancy can be used to improve the reliability and may be required in the both RAN and the CN.
  - Packet duplication can be used in the RAN
  - Redundant paths can be configured in the CN
  - Redundant paths can be correlated via a common control-plane.
- Since redundancy may impact energy and resource efficiency, redundancy should only be used only when required.
- It should be possible to deactivate the redundancy in the RAN and the CN independently.
  - The decision to use duplication depends on the target reliability in the RAN and the CN.
  - Duplication may be required in both the RAN and the CN, in the RAN only, in the CN only or neither the RAN nor the CN.
  - Although the decision to use duplication is independent, the solution to enable the duplication in the RAN and/or CN may not be independent.
- A complete E2E solution covering the RAN and CN should be considered rather than a segmented solution, where the RAN and CN independently provide solutions for redundancy.
- The E2E solution for redundancy should be able to avoid a single point of failure.
- New transport/tunnelling protocols can be considered to improve reliability without impacting latency. The new protocol should support the redundancy requirements in the CN.
- The E2E latency depends on the latency in the UE, RAN, CN and AS as well as on the number of forwarding nodes between the RAN access point and the AS.
  - The processing delays at the UE and the AS are non-3GPP delays
- There is a trade-off in the latency budget between the RAN and the CN.
  - The latency requirements of the RAN can be relaxed by decreasing the latency target in the CN. Conversely, the latency in the CN can be relaxed by decreasing the latency target in the RAN.
- For latency-critical services, it is worth assessing whether there is a significant trade-off in the latency budget between the Non-3GPP processing delays and 3GPP processing delays. In case there is:
  - Improving the capabilities of the UE and the AS by reducing the processing delays can reduce the requirements on the 3GPP delays.
  - Developing leaner applications that minimise the payload requirements on the 3GPP system can reduce the requirements on the 3GPP system.
- In order to satisfy the extreme requirements in mobility scenarios, packet duplication should be supported during the handover procedure.
- If the reliability of the UE and the AS are taken into account, it may be necessary to also duplicate the functions in the UE and the AS.
- Fault management techniques may be required at each node. This includes instantiating multiple instances of the same function on different nodes and duplicating storage.

11 LIST OF ACRONYMS

3GPP	3rd Generation Partnership Project
AL	Access Link
AP	Access Point
AS	Application Server
CN	Core Network
DL	Downlink
E2E	End-to-End
eNB	evolved Node B
gNB	generation Node B
IP	Internet Protocol
LTE	Long Term Evolution
MEC	Multi-access Edge Computing
MgNB	Master gNB
NR	New Radio
PDCP	Packet Data Convergence Protocol
QUIC	Quick UDP Internet Connections
RAN	Radio Access Network
RLF	Radio Link Failure
RRC	Radio Resource Control
SgNB	Secondary gNB
UDP	User Data Protocol
UE	User Equipment
UL	Uplink
URLLC	Ultra-Reliable Low Latency Communication
TCP	Transport Control Protocol

12 BIBLIOGRAPHY

[1] NGMN, “5G Extreme Requirements: Operators' view on fundamental trade-offs,” 28 November 2017. [Online].

[2] NGMN, “5G Extreme Requirements: Radio Access Network Solutions”.

[3] NGMN, “NGMN 5G White Paper,” 2015.

[4] NGMN, “NGMN Perspectives on Vertical Industries and Implications for 5G,” 2016.

[5] 3GPP, “Service Requirements for the 5G System,” *TS 22.261.*

[6] 3GPP, “FS_SMARTER - Massive Internet of Things,” *TR 22.861.*

[7] 3GPP, “FS_SMARTER - Critical Communications,” *TR 22.862.*

[8] 3GPP, “FS_SMARTER - enhanced Mobile Broadband,” *TR 22.863.*

[9] ITU, ITU-R M.[IMT-2020.EVAL], 2017.

[10] ITU, ITU-R M.[IMT-2020.TECH PERF REQ], 2017.

[11] 3GPP, New Radio WI, 2017.

[12] 3GPP, LTE URLLC WI, 2017.

[13] 3GPP, “Study on communication for automation in vertical domains (CAV),” *TS 22.804.*

[14] 3GPP, “Feasibility study on LAN support in 5G,” *TR 22.821.*

[15] 3GPP, “Study on Architecture for Next Generation System,” *TR 23.799.*

[16] A. Hilt, G. Járó and B. I., “Availability Prediction of Telecommunication Application Servers Deployed on Cloud,” *Periodica Polytechnica Electrical Engineering and Computer Science,* vol. 60, no. 1, pp. 72-81, 2016.

[17] O. Salmela, Reliability Assessment of Telecommunications Equipment, Helsinki University of Technology , 2005.

[18] [Online]. Available: https://ece.uwaterloo.ca/~ece477/Lectures/ece477_3.ppt.

[19] “Fibre optic: What are achievable distances with single-mode and multi-mode fibre,” [Online]. Available: https://www.universalnetworks.co.uk/faq/fibre-optic/what-are-achievable-distances-single-mode-vs-multi-mode-fibre.

[20] K. Papagiannaki and e. al., “Measurement and Analysis of Single-Hop Delay on an IP Backbone Network,” *IEEE Journal on Selected Areas of Communications,* vol. 21, no. 6, 2003.

[21] N. Zilberman, M. Grosvenor, D. A. Popescu, N. Manihatty-Bojan, G. Antichi, M. Wojcik and A. W. Moore, “Where Has My Time Gone?,” in *International Conference on Passive and Active Network Measurement*, 2017.

[22] “Transport Layer Security (TLS),” [Online]. Available: https://hpbn.co/transport-layer-security-tls.

[23] Huawei, “Evaluation on Packet Duplication in Multi-connectivity,” in *3GPP WG2 R2-1700172*, 2017.

[24] J. Rao and S. Vrzic, “Packet Duplication for URLLC in 5G: Architectural Enhancements and Performance Analysis,” *IEEE Network,* vol. 32, no. 2, pp. 32-40, 2018.

[25] 3GPP SA WG2, “New SID: Study on enhancement of URLLC supporting in 5GC,” *SP-180118,* 2018.

[26] 3GPP SA WG2, “Study on enhancement of URLLC supporting in 5GC,” *TR 23.725,* 2018.

[27] “R2-1805449 (MCG RLF handling in case of NE-DC (TP to 37.340))”.

[28] Q. De Coninck and O. Bonaventure, “Multipath QUIC: Design and Evaluation,” in *In Proceedings of the 13th International Conference on emerging Networking EXperiments and Technologies (CoNEXT '17), ACM*, Incheon, Republic of Korea, 2017.

[29] 3GPP, “Feasibility study for further advancements for E-UTRA (LTE-Advanced),” *TR 36.912.*

[30] ITU-R M.2135, “Guidelines for evaluation of radio interface technologies for IMT-Advanced,” 2008.

[31] N. A. Johansson, E. Y.-P. Wang, E. Eriksson and M. Hessler, “Radio Access for Ultra-Reliable and Low-Latency 5G Communications,” in *IEEE ICC - Workshop on 5G & Beyond - Enabling Technologies and Applications*, 2015.

[32] 3GPP, “Study on scenarios and requirements for next generation access technologies,” *TR. 38.913.*

[33] 3GPP, “Study on new radio access technology Physical layer aspects,” *TR 38.802.*

[34] S. Parkvall, E. Dahlman, A. Furuskar and M. Frenne, “NR: The New 5G Radio Access Technology,” *IEEE Communications Standards Magazine,* 2017.