Green Technologies for 5/6G Service-Based Architecture

Grant Agreement No	**101096925**
Full Title	**Green Technologies for 5/6G Service-Based Architectures**
Start date	**01/01/2023**
End date	**31/12/2026**
Duration	**36**
Project URL	**https://www.6green.eu**
Coordinator	**CONSORZIO NAZIONALE INTERUNIVERSITARIO PER LE TELECOMUNICAZIONI (CNIT)**

Deliverable D2.1

Use and Business Cases, Design and Technology Requirements, and Architecture Specification

Contractual due date	30/06/2023 (M6)	**Actual submission date**	27/06/2023
Nature	Report	**Dissemination Level**	Public
Lead Beneficiary	TIM		
Responsible Author	Elisa Zimaglia (TIM), Roberto Fantini (TIM)		
Contributions from	Arturo Bellin (ATH), Daniele Munaretto (ATH), Daniele Ronzani (ATH), Borja Otura Garcia (ATOS), Chiara Lombardo (CNIT), Nicole Martinelli (CNIT), Eleonora Borgia (CNR), Raffaele Bruno (CNR), Claudio Cicconetti (CNR), Anastasios Zafeiropoulos (ICCS), Luka Koršič (ININ), Janez Sterle (ININ), Rudolf Sušnik (ININ), Jakob Kämpfer (OCULAVIS), Catalin Brezeanu (ORO), Razvan Mihai (ORO), Marius Iordache (ORO), Carmen Patrascu (ORO), Luis Miguel Contreras Murillo (TID), Alejandro Muñiz Da Costa (TID), Knut Kvale (TNOR), Maurizio De Paola (TIM), Jane Frances Pajo (TNOR), Dimitris Klonidis (UBITECH), Thanos Xirophotos (UBITECH)		

Revision history

Version	Issue Date	Changes	Contributor(s)
v0.1	11/04/2023	Initial version, Section 2 draft, UC1 draft, Network latency draft	Roberto Fantini (TIM), Elisa Zimaglia (TIM), Maurizio De Paola (TIM), Rudolf Susnik (ININ), Chiara Lombardo (CNIT)
v0.2	01/06/2023	First draft ready for internal review	All authors
v0.3	06/06/2023	Second draft ready for internal review with section 3.2 revised and 4.3 still to be completed	All authors
v0.4	14/06/2023	Complete draft for internal review	All authors
v0.5	22/06/2023	Version ready for submission	All authors
v1.0	27/06/2023	Final formatting revision	Riccardo Rapuzzi (CNIT)

Disclaimer

The content of the publication herein is the sole responsibility of the publishers and it does not necessarily represent the views expressed by the European Commission or its services.

While the information contained in the documents is believed to be accurate, the authors(s) or any other participant in the 6Green consortium make no warranty of any kind with regard to this material including, but not limited to the implied warranties of merchantability and fitness for a particular purpose.

Neither the 6Green Consortium nor any of its members, their officers, employees or agents shall be responsible or liable in negligence or otherwise howsoever in respect of any inaccuracy or omission herein.

Without derogating from the generality of the foregoing neither the 6Green Consortium nor any of its members, their officers, employees or agents shall be liable for any direct or indirect or consequential loss or damage caused by or arising from any information advice or inaccuracy or omission herein.

Copyright message

© 6Green Consortium, 2023-2026. This deliverable contains original unpublished work except where clearly indicated otherwise. Acknowledgement of previously published material and of the work of others has been made through appropriate citation, quotation or both. Reproduction is authorised provided the source is acknowledged.

Table of Contents

List of Figures

List of Tables

Glossary of terms and abbreviations used

Abbreviation / Term	Description
3GPP	3rd Generation Partnership Project
5GC	5G Core Network
AF	Application Function
AI	Artificial Intelligence
API	Application Programming Interface
B5G	Beyond 5G
BBU	Baseband Unit
BSF	Binding Support Function
BSS	Business Support System
BSSF	BSS Function
BYOD	Bring Your Own Device
CaaS	Containers as a Service
CapEx	Capital Expenditure,
CCMF	Cloud Continuum Management Function
CI/CD	Continuous Integration / Continuous Delivery
CN	Core Network
CNCF	Cloud Native Computing Foundation
CNF	Containerized Network Function
CPRI	Common Public Radio Interface
CPU	Central Process Unit
DaaS	Desktop as a Service
DAG	Directed Acyclic Graph
DRL	Deep Reinforcement Learning
E2E	End-to-End
eBPF	Extended Berkeley Package Filter
EdgeMF	Edge-cloud Management Function
EGMF	Exposure Governance Management Function
EI	Edge Intelligence
ENI	Experiential Networked Intelligence
ENIF	ENI Function
ETSI	European Telecommunications Standards Institute

Abbreviation / Term	Description
FaaS	Function as a Service
FG	Focus Group
FPGA	Field-Programmable Gate Arrays
GHG	GreenHouse Gas
GPU	Graphics Processing Unit
guacd	guacamole proxy deamon
HTTP	HyperText Transfer Protocol
HTTPS	HTTP Secure
IaaS	Infrastructure as a Service
ICT	Information and Communications Technology
IDAF	Infrastructure Data Analytics Function
IoT	Internet of Things
ITU	International Telecommunication Union
KPI	Key Performance Indicator
KVI	Key Value Indicator
LTE	Long-Term Evolution
MDAF	Management Data Analytics Function
MEC	Multi-access Edge Computing
ML	Machine Learning
ML5G	Machine Learning for Future Networks including 5G
NAO	Network Application Orchestrator
NEF	Network Exposure Function
NF	Network Functions
NFVI	Network Function Virtualization Infrastructure
NIC	Network Interface Card
NR	New Radio
NRF	Network Repository Function
NSDAF	Network Slice Data Analytics Function
NSENIF	Network Slice ENI Function
NSI	Network Slice Instance
NSM	Network Service Mesh
NSMF	Network Slice Management Function
NWDAF	Network Data Analytics Function

Abbreviation / Term	Description
OAM	Operations, Administration and Maintenance
ONAP	Open Network Automation Platform
ONAP	Open Network Automation Platform
OpEx	Operational Expenditure
OSM	Open-Source MANO
OTT	Over-the-Top
PCF	Policy Control Function
PCoIP	PC over IP
PID	Process ID
PSNR	Peak Signal-to-Noise Ratio
QoE	Quality of Experience
QoS	Quality of Service
RAM	Random Access Memory
RAN	Radio Access Network
RAPL	Running Average Power Limit
RAT	Radio Access Technology
RDP	Remote Desktop Protocol
RRH	Remote Radio Head
RTT	Round Trip Time
SA	Stand-Alone
SAN	Storage Area Network
SBA	Service-Based Architecture
SBI	Service-Based Interface
SNS	Smart Network and Service
SoTA	State of The Art
SSH	Secure Shell
TCP	Transmission Control Protocol
UC	Use Case
UE	User Equipment
ULL	Ultra-Low Latency
UPF	User Plane Function
UWB	Ultra-Wide Band
vAPP	Vertical Application

Abbreviation / Term	Description
vDesktop	virtual Desktop
VDI	Virtual Desktop Infrastructure
VIM	Virtualized Infrastructure Manager
VNC	Virtual Network Computing
VNF	Virtualized Network Function
WIM	Wide-area Infrastructure Manager
WP	Work Package
XaaS	Anything as a Service
XR	eXtended Reality
ZSM	Zero-touch Service Management

Executive Summary

This deliverable is the first one released by the 6Green project. It provides a summary of the project's objectives and vision, describes the main technological enablers and the proposed service-based architecture, and discusses the selected use cases, including their underlying architecture, the associated business framework, the relevant stakeholders involved in their execution, and their Key Performance Indicators (KPIs).

The project aims to create an innovative, service-based, and comprehensive ecosystem that promotes energy efficiency across the entire 5/6G value chain. The ultimate objective is to reduce the carbon footprint of 5/6G networks and vertical applications. To achieve this target, the project will:

1 Transition from traditional "Cloud Agility" to "Edge Agility" by seamlessly moving applications and services across the edge-cloud continuum in real-time. This will be accomplished by utilizing extended service mesh technologies and efficient traffic steering mechanisms.
2 Introduce "Green Elasticity" to automate the provisioning and optimization of network and computing resources. This includes dynamic task transfers and AI-based decision engines for energy consumption optimization.
3 Enable "Carbon/Energy-aware Backpressure" by monitoring energy metrics, defining energy/carbon-aware KPIs, and autonomously learning power consumption profiles.
4 Integrate the 5/6G Service-Based Architecture (SBA) with green AI mechanisms to reduce energy consumption, leverage renewable energy, and optimize coordination with stakeholders and vertical applications (vApps).
5 Establish new green business models and Decarbonized Service Agreements. It will introduce self-service slice Business Support System (BSS), allowing dynamic agreement renegotiation, and incentivizing sustainable policies.

As described in the document, the primary enabling technologies that will facilitate the realization of the project's vision will be Artificial Intelligence, Cloud-native approaches, and the Edge-cloud continuum. These technologies will be integrated into an SBA composed of different frameworks that can interact synergistically to achieve the project's decarbonization target.

To demonstrate and validate project achievements three use cases have been selected:

- Use Case 1 (UC1): "Critical Operation Maintenance during Energy-Constraint Disaster Scenarios". This use case focuses on critical infrastructure operations and the need to maintain functionality even in the face of disruptions. It assumes distributed computing capabilities, including a non-public 5G-Advanced/6G network. The objective is to prioritize and relocate functions to minimize consequences when certain parts of the computing capabilities are unavailable.
- Use Case 2 (UC2): "Energy-Efficient Augmented Reality Remote Assistance System". This use case aims to reduce carbon emissions and machine downtimes by replacing business trips with remote collaboration tools. It enables experts to guide on-site technicians using smart glasses or smartphones/tablets, improving resource usage, and lowering operational costs. The use case aims to enhance the Quality of Experience (QoE) for end users while ensuring energy-efficient algorithms and adapting the Quality of Service (QoS) by considering server location changes and resource usage improvements. It aligns with the goals of the 6Green project to enhance flexibility, scalability, and sustainability in the ecosystem.
- Use Case 3 (UC3): "Zero-Carbon Clientless Virtual Enterprise Desktop as-a-Service". In this scenario, end users connected to a 5G mobile network can access remote desktop servers deployed either in a central cloud or in a public cloud. With the advancements of 5/6G technology, Desktop as a Service (DaaS) solutions can benefit from low latency and shift computational intelligence to servers, reducing operational costs and greenhouse gas emissions.

These use cases will allow to evaluate the 6Green platform's performance, conduct live demonstrations, and disseminate outcomes to industry and scientific communities.

1 Introduction

This document is the first deliverable of the 6Green Project and represents the initial activities carried out in Work Package 2 (WP2), “Green Enabling Technologies for Cloud-Native Services.” The goal of the 6Green project is to create an innovative, service-based, and comprehensive ecosystem that expands the communication infrastructure into a sustainable, interconnected, and greener end-to-end inter-computing system. It aims to promote energy efficiency across the entire 5/6G value chain and reduce the carbon footprint of 5/6G networks and vertical applications.

The 6Green project is structured around three main research axes, which correspond to three administrative domains/layers within the architecture: the **5/6G Edge-Cloud Infrastructure**, the **Network Platform** and the **Vertical Application** domains. The corresponding research areas are referred to as “Enabling Technologies for Cloud-Native Service Meshes,” “the 6Green Service-based Architecture,” and “Vertical Application Orchestration within the 5/6G Green Economy.” These axes must closely collaborate to implement the holistic vision of 6Green.

WP2 primarily focuses on the activities related to the first research axis of the project: “Green Enabling Technologies for Cloud-Native Services.” Within WP2, the initial activities involved refining the architectural definition of the 6Green ecosystem and validating use cases for the technologies and solutions developed by the project. This includes identifying the roles of different stakeholders, determining system and use cases requirements, and establishing key performance indicators for the identified use cases.

With this objective in mind, this document is organized as follows. The first section introduces the deliverable and outlines the overall structure of the document. Section 2 delves into the project's vision and ambitions, elaborating on the main objectives and introducing the key innovations: **Edge Agility, Green Elasticity, and Energy-aware Backpressure**. These innovations form the foundation of the 6Green vision.

Section 3 presents the primary enabling technologies that will facilitate the realization of the project's vision, including Artificial Intelligence, Cloud-native approaches, and the Edge-cloud continuum. Additionally, it discusses 6G requirements and describes the proposed 6Green SBA, exploring the various frameworks within the architecture. The section emphasizes how these frameworks can interact synergistically to achieve the project's decarbonization target.

Section 4 is dedicated to the three selected use cases in the project: “Critical Operation Maintenance during Energy-Constraint Disaster Scenarios,” “Energy-Efficient Augmented Reality Remote Assistance System,” and “Zero-Carbon Clientless Virtual Enterprise Desktop as-a-Service.”" For each use case, a general description is provided, along with a discussion on the roles of different stakeholders. Furthermore, an overview of relevant technical enablers specific to each use case is presented, followed by a list of required functional enablers and key performance indicators.

Finally, Section 5 concludes the document by summarizing the main elements discussed throughout the deliverable and highlighting their significance in the upcoming project activities.

2 Project Vision and Ambitions

The Fifth-Generation (5G) of radio mobile networks and edge computing technologies are transforming the cloud into a flexible communication and inter-compute continuum. Recent studies indicate that 5G and edge computing will cause a noticeable increase in computing resources, increasing the associated infrastructure OpEx and CapEx, as well as their carbon footprint and energy requirements. To cope with this problem and meet sustainable growth targets, the 5/6G continuum needs to evolve new foundation paradigms specifically addressing energy and carbon footprints of the overall ecosystem.

The 6Green project aims to create an innovative service-based and holistic ecosystem to extend 5/6G networks and vertical applications, reducing their carbon footprint by a factor of 10 or more. It will exploit and extend state-of-the-art cloud-native technologies and the Beyond 5G (B5G) Service-Based Architecture (SBA) with new cross-domain enablers to boost the global ecosystem flexibility, scalability and sustainability.

With this objective in mind the project aims to:

- Transition from classical “Cloud Agility” schemes to a novel “Edge Agility” paradigm. Edge Agility is a paradigm that aims to transition from the traditional "Cloud Agility" approach to a new concept of dynamically moving applications and services within the edge-cloud continuum. In this new paradigm, vertical applications and network services will be deployed and scaled based on workload requirements, allowing seamless relocation across different geographical areas. This will be achieved by extending cloud-native service mesh technologies with new capabilities, such as smart scaling and efficient traffic steering at the microservice level.

  The project focuses on creating a holistic and cross-domain ecosystem that enables the smooth movement of Network Functions (NFs) and vertical Applications (vApps) within the edge-cloud continuum. It involves introducing novel features at the service mesh sidecars [JM20], including the ability to scale NFs and vApps to zero when not in use, and providing mechanisms for proactive and reactive relocation operations based on policies. The energy and carbon-aware cost of relocation, as well as the computational cost of automatic decisions, will be considered.

  Edge Agility will play a crucial role in providing intelligent and automated horizontal scalability for vertical applications and their slices across the 5/6G edge-cloud continuum. It will enable the redistribution of workload in response to user or infrastructure events, and support seamless operations, e.g., during UE handovers or policy-driven events. By integrating this capability with the control plane, it will allow for adjustment of the latency budget between connectivity and computing.

  Management operations will be conducted to quickly scale to zero the footprint of slices and vertical applications in unused areas of the continuum, while efficiently resuming operation when needed. The 5/6G Application Function will interact bidirectionally with Edge Agility, coordinating and synchronizing network and application-level orchestration. Overall, Edge Agility aims to provide a smart, fast, and automated approach to handle the dynamic requirements of applications and services in the edge-cloud continuum.

- Introduce the “Green Elasticity” paradigm. Green Elasticity is a paradigm that aims to introduce joint, automatic, and rapid provisioning, adaptation, and de-provisioning of network and edge computing resources in the context of 5/6G. It involves extending cloud-native service mesh technologies with new capabilities to dynamically transfer critical tasks between software and hardware accelerators. Automated decision engines will be developed to manage and trade off various types of compute and network resources, with the goal of meeting service requirements while reducing overall energy consumption.

The 6Green project focuses on efficiently leveraging new compute resources to offload network and computing critical tasks of 5/6G Network Functions (NFs) and vertical Applications (vApps). The improved SBA architecture of 6Green will serve as the energy-efficient foundation for the physically distributed data centers within the 5/6G edge-cloud continuum, as well as the connected vApps. Communication efficiency between these elements and the joint optimization of the overall energy consumption will be key considerations for service decisions.

Green Elasticity aims to provide energy-aware, hardware-assisted acceleration to NFs and vApps, enabling smart vertical scalability across the three domains of 5/6G environments: the vertical domain, the network slice, and the underlying infrastructure. Hardware acceleration, such as programmable hardware and GPUs or FPGAs, can significantly reduce processing latency compared to pure software. The energy efficiency of hardware acceleration depends on the volume of workload offloaded from the software level. It becomes advantageous when applied to large volumes of time-varying workloads or when optimizing end-to-end configurations.

The paradigm of Green Elasticity allows for dynamic distribution of time-varying, end-to-end latency budgets for vertical applications across domains while optimizing the trade-off between energy/carbon footprint and network/application performance. Elasticity operations will no longer be constrained to a single technological domain but will propagate to other domains, enabling the adaptation and consolidation of service meshes for optimal utilization of available computing and networking resources.

- Enable the "Carbon/Energy-aware Backpressure" in the 5/6G SBA. The 6Green project aims to enable Carbon/Energy-aware Backpressure in the 5/6G Service-Based Architecture (SBA), allowing observation of energy consumption and carbon footprint metrics at the geographical infrastructure level. This capability will be decomposed and mapped on a per-network slice and per-vertical application basis. To achieve this, 6Green will define energy- and carbon-aware Key Performance Indicators (KPIs) to be exposed to each stakeholder domain. New monitoring and analytics tools will be provided to the cloud-native infrastructure, enabling data analytics for mining and estimating past and future footprint induced by slices and vApps. Additionally, the project platform will autonomously learn power consumption profiling of different slice and vApp categories. The Energy-aware Backpressure encompasses cross-domain observability mechanisms and analytics to evaluate the energy and carbon footprint induced by vertical applications, slices, and the overall 5/6G network on the edge-cloud infrastructure. It aims to process, infer, and expose this information at both the 5/6G SBA and vertical application levels, including their network slices. Hardware-level energy consumption metrics will be collected, considering renewable energy contributions, and mapped to each hosted tenant using adaptive AI-driven analytics. The SBA will be extended to acquire these energy consumption metrics and classify them, exposing them to the accountable vertical. The backpressure metrics will be jointly consumed by optimization engines operating in all stakeholder domains through a cooperative win-win approach.
- Integrate the B5/6G SBA with new green AI mechanisms, to dynamically and holistically reduce the energy consumption/carbon footprint of both the network and the vertical applications, as well as exploiting renewable energy sources supplying the distributed infrastructure. To achieve this, 6Green will conceive and implement novel AI-based decision engines to flexibly apply/drive Edge Agility and Green Elasticity operations over the overall network or single slices. It will bound resources exposed to vApps with availability of 5/6G network-level and infrastructure-level resources, and the availability of renewable energy sources. It will support multi-objective decision policies considering the trade-off between performance and the impact on energy requirements/induced greenhouse

gas emissions. 6Green will design AI-mechanisms as a distributed, modular, and hierarchical swarm, able to divide and conquer the cross-domain environments, and to harmonize different stakeholders' policies and objective at the 5/6G platform level. Additionally, 6Green will extend 3GPP specifications towards the observability of cross-domain metrics, and to drive in-network and in-vApp intelligence to zero-touch energy efficient reconfigurations. Finally, 6Green will evolve the SBA intelligence to act on 5/6G control and management planes in a fully integrated fashion by design, as well as coordinate with vApps hosted in the edge-cloud continuum.

- Enable new green 5/6G business models and Decarbonized Service Agreements among stakeholders. It involves extending the 5/6G SBA to provide vertical stakeholders with self-service slice Business Support System (BSS), supporting dynamic renegotiation of service level agreements, providing incentives for embracing sustainable policies, drawing new win-win business models based on green economy, and allowing vertical stakeholders to establish business models and service level agreements to assess their zero-carbon impact on the 5/6G network, while cutting their greenhouse gas emissions.
- Demonstrate and validate the project achievements in three use-cases. To achieve this, the 6Green platform must be realized and integrated into the project testbed. Three future-proof Use-Cases, which will be presented in Section 0, will be evolved and integrated in the platform, and they will allow to experimentally evaluate and assess the performance of the 6Green prototypes, organize live demonstrations in public events, and disseminate the project outcomes to the industrial and scientific communities. The 6Green platform is designed to target sustainability goals while strengthening 5/6G technologies towards novel services requiring strict timing, fast mobility, and highly dynamic situations.

These innovations form the cornerstone of the 6Green vision, enabling energy- and carbon-efficient configurations as well as the assessment of indirect energy/carbon footprints. They encompass a range of research activities and technology advancements that can be categorized into three fields aligned with the project's vision and approach: **5/6G Sustainability**, **5/6G SBA Evolution**, and **5/6G AI**.

The main 6Green ambition is to evolve the 5/6G system and the overall cloud continuum to reach a full **5/6G sustainability**, not bound to the network/computing capacity anymore, but rather to the real usage of resources from an energy consumption perspective. This will involve extending the 5G SBA to transform network slices into dynamic environments able to coordinate network- and application-side workloads and requirements and enable them to jointly target the reduction of their energy and carbon footprints. This will be achieved through a cognitive and holistic approach, making overlying systems and applications aware of underlying network green reconfigurations. The 6Green project aims to reduce energy/carbon wastes by reducing resource/energy usage in actively used facilities and zeroing consumption in all areas declared in the network slice but not currently used. It will take advantage of state-of-the-art works on energy efficiency for last generation cloud-native systems, and propose new business model-based green economy best practises to enable MNOs and Vertical Stakeholders to reduce the overall energy footprint and reach a Decarbonization Service Agreement.

6Green will contribute to significantly **evolve the 5G SBA** under different aspects that will lead to the introduction of novel functions and the extension of already existing services. The first evolution line will involve a strengthened support to 5G control and management observability, energy-aware backpressure, and relevant metrics to drive the agility and elasticity mechanisms. The SBA exposure layer towards vertical Application Functions (AFs) will be notably extended to support the aforementioned green business models. A further key ambition will be to support the green holistic approach by including computing

resources/requirements of the vertical application within the slice (re)negotiation. To support edge agility and green elasticity, the 6Green SBA will also target the full integration of 5/6G control and management-plane functions. This will allow not only to "scale-to-zero" slice functions in non-used areas and to rapidly scale up according to the workload, but even a better integration with power management mechanisms offered by the bare-metal edge-cloud infrastructure.

The 6Green ambition in this area includes the native integration of Artificial Intelligence engines for zero-touch SBA automation at both overall-network and single-slice levels. These engines will act upon the events and analytics streams coming from both the infrastructure and the vertical applications in order to rapidly reach energy efficient configurations.

The 6Green ambition on **5/6G Artificial Intelligence (AI)** is double-fold and cross-connected with the ambitions on sustainability and 5/6G SBA evolution. AI will play a central role within the project since it will be the key means to estimate the energy and carbon footprint that vertical applications and network slices are inducing into cloud-edge infrastructure. Exploitation of capabilities offered by beyond-5G networks is key to move towards this direction. 3GPP has specified a set of northbound APIs [23.971] for managing network functions or making available aggregated views of data to Over-the-Top (OTT) players. The ETSI ENI (Experiential Networked Intelligence Industry Specification Group) [ENI005] defines a Cognitive Network Management architecture that adopts AI and context-aware policies to adjust offered services based on changes in user needs, environmental conditions, and business goals. Furthermore, the ITU FG ML5G (Focus Group on Machine Learning for Future Networks including 5G) [ITUML5G] drafted ten technical specifications for ML for future networks, including interfaces, architectures, protocols, algorithms and data formats. 6Green is going to build upon these specifications and provide interoperable and open APIs for the ease integration and execution of data analytics processes within orchestration mechanisms.

6Green aims to develop mechanisms and techniques for both real-time and long-term analytics to provide accurate and efficient analytics technologies specific for the dynamic deployment and lifecycle management of vertical applications and network slices. Analytics will be performed on both real-time and stored profiling data to identify patterns that can help optimize the behaviour of both 5/6G network and applications within the edge cloud-continuum.

Zero-touch decision engines will be developed to enable AI swarms to jointly act on 5/6G control and management planes, and on vertical application deployments over the edge-cloud continuum.

The ambition is to support a wide spectrum of "plug-and-play" policies targeting diverse and multiple objectives, such as power, energy and carbon capping, and zero-carbon operations.

The 6Green platform is designed to account for and exploit the availability of renewable energy sources within the edge-cloud continuum. It also considers the impact of AI mechanisms as an integral part of the energy and carbon footprint associated to the overall network or to single slices/vertical applications. The project ambition is not only to contribute and extend Green AI research, but also to enable the 6Green SBA control and management planes to carefully ponder the execution of computationally hungry algorithms against the potential energy savings that they can bring to the infrastructure, for optimally balancing potential savings coming from AI decisions and the AI resource footprint itself.

3 Enabling Technologies and Service-Based Architecture (SBA)

This section examines the essential technologies that will contribute to achieving the objectives of the 6Green project. These technologies include Artificial Intelligence, the Cloud-native paradigm, and the concept of the edge-cloud continuum. Additionally, since 6G will play a crucial role in the 6Green platform, Section 3.2 provides a summary of the most significant 6G requirements for the project. Lastly, Section 3.3 presents the proposed evolution for the service-based architecture, which integrates all the technological elements described in this section to accomplish the project's decarbonization goal.

3.1 Enabling Technologies

3.1.1 Artificial Intelligence

Among the technologies considered pivotal for moving beyond 5G, Artificial Intelligence is definitely the most distinctive. Many publications have appeared, especially in the last several years, that highlight the contribution of AI to 5/6G.

Both the academia and industry are working for the future research in order to optimize the 5G and 6G network. As highlighted in [KNV+18], "*The next-generation wireless networks are evolving into very complex systems because of the very diversified service requirements, heterogeneity in applications, devices, and networks*". This means that the network operators need to keep up with all these requirements, together with the extension of coverage with scarce resources and limited capital. In [YAX+20] the authors introduce an integrated space-air-ground-underwater network for the future 6G, in order to have seamless and near instant super-connectivity. Given all this, both authors in [KVV+18, YAX+20] emphasizes the need of automation by using advance data analytics and Artificial Intelligence for the next wireless generation, in order to facilitate network planning, control, and optimization as well as advance energy management strategies, which, if done manually, will be more time consuming and susceptible to human error. Artificial intelligence can also be implemented in Edge-Intelligence (EI), which will enable edge equipment to provide real-time responses and perform model training and local inference, without entirely relying on the cloud. Authors in [JSZ+21] discuss the prior knowledge on this topic, while exposing two typical case studies and discussing potential challenges in typical embodiments scenarios of EI.

When talking about AI in the context of the SBA, the Network Data Analytics Function (NWDAF) is usually part of the discussion. The NWDAF has first been introduced in 3GPP Release-15 specifications as part of Core (5GC) for supporting intelligent and autonomous network operations and service management [NZS+22]. The 5GCore can deploy multiple instances of NWDAF: they do not need to provide the same type of analytics, since some of them can be specialized in providing only some types of analytics. An analytics ID information element identifies the type of analytics that an NWDAF instance can generate, and a single NWDAF can provide multiple ids. A prototype of the NWDAF has been proposed by [CMS22], where the authors analysed the network generated by exploring the interaction between core NF-NF using machine learning techniques, in particular unsupervised learning to determine similarities among them. Another case study is the analysis of NWDAF-collected core network function data from an Open5GS and UERANSIM implementation, by considering the 5G core messages between the Binding Support Function (BSF) and the Network Repository function (NRF) [CMS22b]. In general, the NWDAF is a system module in the 5GC architecture that provides network data collection and analytics, thus enabling closed-loop optimization, network automation and service orchestration. It has been designed to acquire data from 5GC NFs and expose analytic services on the

south bound for seamless integration with the rest of the NFs and the operator's cloud environment [23.288]. As such, it allows various data-driven AI/Machine Learning (ML)-based analytics technologies to be integrated with 5G networks.

The 6Green ambition on AI goes beyond observability and analytics, aiming to heavily rely on these techniques for automated decisions on the control and management operations towards the infrastructure, the SBA NFs and the slices. To this end, the ETSI Experiential Networked Intelligence (ENI) is the most mature available solution to embed AI in the network. The authors in [EGD19] use the concept of resource elasticity and ENI architecture as a base to coordinate the elastic slice lifecycle management, proving it by showing the applicability of AI in three use case of different management and orchestration problems where elasticity can be exploited. ETSI ENI defines standards for cognitive network management and defines a general architecture that may be applied to virtually all management aspects in future such as infrastructure management, network operation, service orchestration and management, and network assurance, for which it provides a standardized approach for intelligently managing network slice instances (NSIs) in a closed-loop system, enabling the automation and AI algorithms to dynamically configure NSIs in a scalable and flexible manner [WRE20]. The ETSI Zero-touch Service Management (ZSM) goal is to accelerate the definition of the required end-to-end architecture and solutions, while enabling an autonomous network system capable of self-configuration, self-monitoring, self-healing, and self-optimization based on service-level policies and rules without human intervention. In compliance to the later, the authors in [CKB22] introduce a novel distributed management and orchestration framework that addresses the challenge of handling a massive number of network slices as envisioned in 6G while using AI-drive closed loop controls.

3.1.2 Cloud-Native

Cloud-based infrastructure for 5G and future 6G networks is all about virtualization and disaggregating NFs from restrictive hardware. The development of the 5G network itself is based on microservice and Cloud-native approaches, that breaks down the system into smaller parts that are individually deployable and are interconnected in a cloud computing environment. This architecture is complemented with edge cloud computing to further increase reliability by introducing real-time data processing.

Implementing cloud-based infrastructure for 5G and 6G networks does come with a couple of challenges. For instance, multi-domain orchestration is necessary to synchronize and oversee network services across various Virtualized Infrastructure Managers (VIMs) or clouds. This enables support for diverse use cases, such as having the data plane in one edge cloud while the control plane resides in a public cloud. While OpenStack has been widely adopted for multi-cloud environments, Kubernetes is gaining popularity due to its ability to deliver consistent orchestration across different clouds.

The influence of hyperscale public cloud providers is an important factor to consider. The support for containers in open-source communities like Open Network Automation Platform (ONAP) [ONA17] and Open-Source MANO (OSM) [OSM16] highlights the increasing impact of these cloud providers on the telecommunications industry. Looking ahead, one question arises: can a mature container orchestrator engine like Kubernetes become a full-fledged VNF orchestrator? Kubernetes already provides diverse networking options, and projects like Network Service Mesh (NSM) within the Cloud Native Computing Foundation (CNCF) [CNF17] are emerging to support a wide range of use cases. Moreover, Kubernetes' scalability allows it to be deployed at the network edge using lightweight versions like K8s, enabling the deployment of 5G services that rely on multi-access Edge Computing (MEC) deployments.

However, achieving a fully Cloud-Native Telco system still faces certain barriers. Firstly, the relative newness of Cloud-Native compared to other approaches and technologies poses a challenge. Established non-Cloud Native solutions have been available for a long time and offer shorter time-to-market, making them more

appealing from a business perspective. Another obstacle is the lack of strict standards in Cloud-Native technologies. While standards like ETSI GS NFV-IFA 013 [IFA013] provide guidance for NFV architectures, Cloud-Native predominantly relies on open-source software, which offers more adaptability but less standardized practices.

Nevertheless, the demand for Cloud-Native solutions in the telecommunications industry continues to grow. The technology is maturing, and solutions for Cloud-Native telco core functions and applications are emerging. Standardization efforts, along with the adoption of frameworks like the Open API framework, are expected to address compliance and interoperability concerns, ensuring that the benefits of Cloud-Native infrastructure are fully realized in the telecommunications domain.

In [WBB+20] the authors highlight the three features a deployment technology required to deploy arbitrary cloud-native applications by analyzing the current research around what cloud-native means, thus "*performing a first step towards classifying and assessing the support for deploying cloud-native applications*". To ensure independent deployment, every microservice is developed, packaged, and released using an autonomous, isolated unit of environment. A first solution that [NMM+16] proposes is to deploy a service per physical servers or virtual machines. The problem that arises is the cost, in fact "*even in the "cheapest" cloud-hosting environment the budget for a setup utilizing thousands of servers would be significantly high, probably higher than what most companies can afford or would like to spend*." So, the second solution it proposes is the use of containers, which "*provide a modern isolation solution with practically zero overhead*", since hundreds of containers can be run on a single host.

Although all microservice approaches have a series of benefits in a cloud/edge environment, such as isolating different components to be resilient in a modularized microservice architecture and reduction of the overall bandwidth occupation, they are still quite complex to manage. [WGN+19] highlights that user mobility can result in frequent switch of nearby edge clouds, which increases the service delay when users move away from their serving edge clouds.

One of the ways for better orchestration between the different modules of the microservice architecture is the usage of sidecar-proxies for control and monitoring outside of the service. These are called Service Mesh. In [ARZ+22] the authors compare different service mesh platforms to the performance of each architecture and evaluate their RAM usage and latency. Van-Binh and Duong Younghan Kim in [DK23] proposed a holistic 5G Core Service Mesh architecture to enhance security for 5G core networks using Istio and Cilium. As mentioned before, microservice architectures use container technologies for deployment, and Kubernetes is the de facto tool to orchestrate the container-based architecture. In [WML+22] the authors introduce a microservice monitoring and analysis system called KMamiz, which is able to monitor and analyse system features even as the system serves a heavy traffic load.

Another aspect is the overall resource scheduling in cloud-native environments. The resources are provisioned based on the QoS requirements of the application, but there are some difficulties in the selection of the appropriate resource scheduling method for workloads in order increase the effectiveness of resource utilization. This is because customers and providers having different aims, for instance one wants best execution time with minimal cost, while the other wants to maximize utilization with minimal investment. This causes a lack of communication with one with the other, thus making resource scheduling more of a challenge. [SC16] highlights this obstacle and describes the review technique used to find and to analyse the available existing research, research questions and searching criteria.

The aim of our work is to take a step further, by allowing monitoring of service mesh architectures in 5/6G cloud-native infrastructures by optimizing the execution environment of microservices and their mesh interconnectivity and network-related functions. It will monitor power and resource metrics from the

hardware components of the servers (i.e., CPUs, memory, disks, network I/O) to the hosted containers/pods. To do so, we will user monitoring agents such as Kepler [KEP22] and Scaphandre [SCA20].

Kepler is open-source tool founded by Red Hat in collaboration with IBM Research, which captures the power usage metrics from a Kubernetes cluster. In particular, the CPU, GPU and RAM power consumption information is collected using the extended Berkeley Package Filter (eBPF) in the Linux kernel and RAPL (Running Average Power Limit) metrics, and data can be exported as a set of Prometheus metrics and displayed in dashboards using Graphana. One of the aims of Kepler is the integration of carbon intensity data [LS22]-[FOS23].

Scaphandre is another open-source metrology agent that shows the power consumption of a single process. It has been developed using RUST. It uses the PowercapRAPL sensor, which reads the values of the energy counters from powercap. It then stores those values, and does the same for the CPU usage statistics and for each running process on the machine at that time. From there, it is possible to compute the ratio of CPU time actively spent for a given PID (Process ID) over the CPU time actively spent doing something, in order to estimate the subset of power consumption that is related to that PID on a given timeframe. As in Kepler, and data can be exported as a set of Prometheus metrics and displayed in dashboards using Graphana [SCA20b].

In particular, we will focus on Scaphandre, since it can measure power consumption on bare-metal hosts or virtual machines from the hosts, as well as exposing power consumption metrics of a virtual machine, to allow manipulating those metrics in the VM as if it was a bare metal machine, whereas Kepler exposes only power consumption metrics of Kubernetes, which is quite constrictive for our aims.

3.1.3 Edge-Cloud Continuum

Over the past few years, cloud and edge computing technologies have rapidly gained prominence as the primary computing paradigms. They cater to a wide range of application domains with varying Quality of Service (QoS) demands. Traditional cloud and edge computing systems have introduced numerous orchestration solutions to address specific infrastructure aspects (such as cloud-native computing systems and IoT device management) as well as application requirements (such as ultra-low latency and high computational intensity) [TDM20] [Cos22].

The current centralized computing model is being expanded to accommodate the needs of applications deployed across different parts of the computing infrastructure. This expansion is taking the form of distributed computing continuum systems [DCD22]. By computing continuum, we mean the integration of computing and communication resources and services across an end-to-end infrastructure [Ros22] [Kim21]. The computing continuum expands the existing cloud computing infrastructure by incorporating edge computing and Internet of Things (IoT) computing devices. The goal is to facilitate the deployment and management of distributed applications throughout the continuum while ensuring energy efficiency, high performance, and meeting security and privacy requirements.

The shift towards distributed computing continuum systems presents several challenges that need to be addressed. One of these challenges is the requirement to approach management aspects from a holistic standpoint, where different systems must collaborate towards a shared objective. Operating within the computing continuum involves complex systems, as applications leverage multiple computing tiers and orchestration stacks to function effectively. Within the computing continuum framework, the challenges encountered necessitate the distribution of orchestration across distributed cloud environments, with the aim of approaching management from a comprehensive standpoint. Each stakeholder within the computing continuum, whether they are application, network, infrastructure providers, or end users, establishes their own objectives for application operation. However, a crucial question arises regarding how these objectives can be accurately measured or enforced. Moreover, the computing continuum encompasses a multitude of

resources spread across different computing tiers and layers of abstraction. Ensuring the fulfilment of objectives while maintaining system equilibrium presents another unresolved inquiry.

Several studies have explored orchestration mechanisms for the computing continuum, providing insights into collaborative orchestration actions [Fu22] [Smi21] [Kok22]. These actions encompass managing the lifecycle of applications (e.g., creation, self-healing, microservice restart), resource allocation (compute offloading, scaling, live migration), software-defined network management (e.g., mobility management), and coordination of monitoring mechanisms. The combination of autonomy and distributed intelligence among orchestration components is considered crucial for effective orchestration mechanisms. Achieving this requires leveraging techniques from multi-agent systems management and the joint application of AI techniques [Ros22] [Kok22]. Towards this direction, the development of new orchestration paradigms that effectively balance local autonomy with centralized control is considered. Although local autonomy offers various advantages, it is not enough on its own to guarantee the achievement of objectives related to distributed application deployments and multi-cloud infrastructures. This is mainly because there is a difficulty in attaining system-wide goals when agents operate in complete autonomy or even when they collaborate only within local contexts.

Serverless is a mature cloud technology that simplifies the development of complex applications through its programming model called Function-as-a-Service (FaaS) [Sch21]. The model allows a developer to break down the logic of the application into stateless elementary functions that can be composed in chains or more complex structures, such as Directed Acyclic Graphs (DAGs) [Bha21], which are executed in containers. Since the output of a function invocation depends only on the arguments, it is possible for the service provider to adapt the number of instances currently active for a given function to follow closely the instantaneous demands. At times, while there is absolutely no call of a given function, the number of corresponding containers can be even reduced to zero, which can be an effective measure to reduce the energy consumed by back-end services [Pat21], compared to microservice architectures where a container is typically required to remain alive even if temporarily unused. A typical deployment of a serverless platform is illustrated in Figure 1, where two clients access remote services provided in the cloud by a scalable container infrastructure managed by an orchestrator, which decides how many instances of each function (*f* and *g*, in the example) should be deployed over time. All the function requests are received by the same entry point, which dispatches them to the respective function executors. Since real applications are very often stateful, i.e., the execution of the logic also depends on some state associated to the user/session [Jin21], such state is hosted on an external storage service or in-memory database so that each individual function execution remains stateless from the point of the view of the

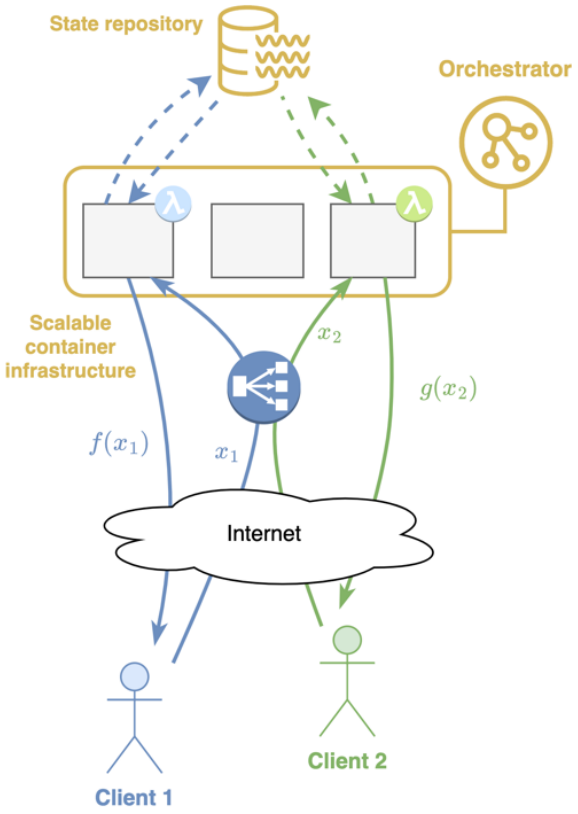

Figure 1: Typical deployment of a serverless platform.

orchestrator. This works well enough in many cloud-based use cases[1] but the same approach cannot be adopted "as-is" at the edge, especially for data-intensive applications such as ML/AI training [Rau21]: the advantages of keeping processing closer to the clients can be nullified by the overhead, in terms of latency and bandwidth, of accessing a remote service to read/update the application's state at each function invocation, as illustrated in Figure 2.

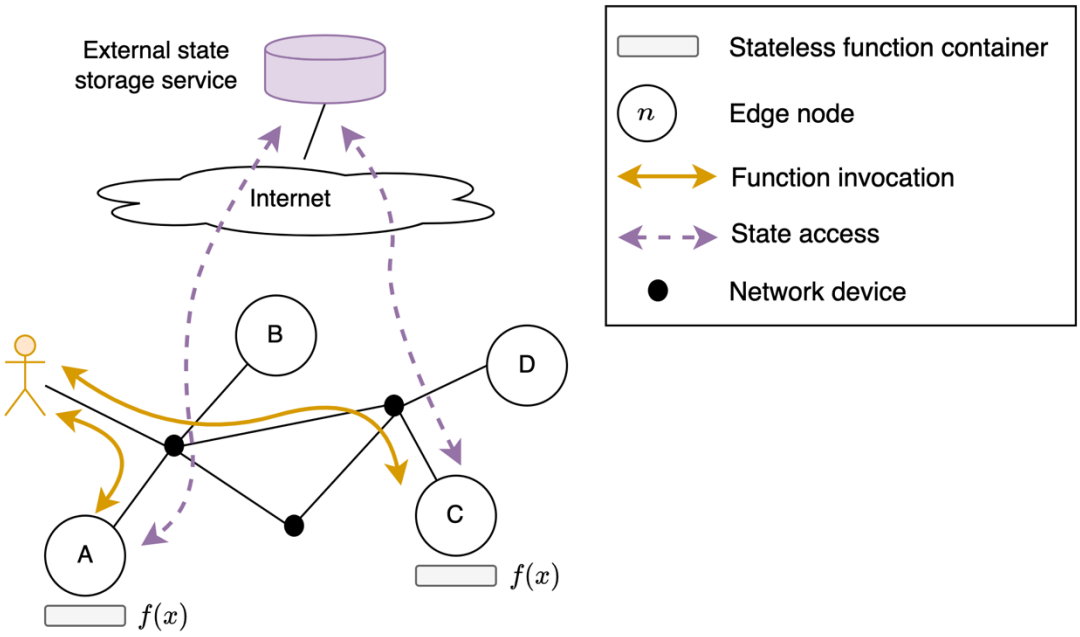

Figure 2: Read/update process of application's state at edge nodes.

In the literature this problem has been recently studied [Sre20, Heu21, Cic22a, Cic22b] but a conclusive solution has not been found yet. Another research direction that requires further investigation is the effect of cold-start on the performance: this phenomenon refers to the extra time needed by the container infrastructure to spawn a new function executor, which includes the setup of a new container instance, the loading of the run-time environment (e.g., Node.js or Python), and the configuration of the network mesh that interconnects the container with the rest of the processing and communication infrastructure. Due to cold-start, the latency of FaaS invocations is often considered unpredictable, with large variations between typical and high-tail values [Man18]. At the edge, cold-start effects are exacerbated by the more limited computational capabilities of the servers and their geographical distribution [Tze21]. Possible solutions to mitigate the cold-start effects, and thus make the latency performance of FaaS more predictable and stable, include caching [Che23] and snapshotting [Kat22], but research in both areas is still in its infancy.

[1] This is not true for *all use cases*. In March 2023 Prime Video announced in a famous blog post the choice to shift their media conversion service from a serverless architecture (using AWS Lambda) to a monolith solution (using Amazon ECS), mainly due to the overhead caused by stateful data access in some of the internal processing functions. The move is claimed to save 90% of the operational costs and the announcement caused widespread discussions in the scientific and industrial research cloud computing community.
Blog post: https://www.primevideotech.com/video-streaming/scaling-up-the-prime-video-audio-video-monitoring-service-and-reducing-costs-by-90

3.2 6G Requirements

6G is currently in the incipient development phase and the requirements discussed in this section are identified in relation with the general envisioned 6G use cases, as identified by various sources, such as [NGMN23] and [PBC+20].

6G provides several network characteristics, as Intelligent Networks, Flexible Networks, Sustainable and Secure Network, Sensing and Localization Network and Management and Orchestration for the Programable Networks, used as inputs for the targeted requirements.

The novel requirements of 6G should be analysed and identified based on 6G use cases and vertical's specificity, introducing the novel 6G UCs family characteristics, as well as new network and services deployments methods aligned with the environmental condition.

6G is enabling:

- services delivered by the next telco networks,
- CaaS implementation of the network functions and software components,
- cloud-native implementation,
- intent-based networks,

that are characterized by several attributes, as for example AI-aaS, network high-resolution sensing services for various application or high-accuracy positioning, hyperconnected networks, high-resilient, scalable and secured networks.

New sets of KPIs or KVIs (Key Value Indicators) and new services QoS should be implemented by the 6G networks, based on different architectural principles or capabilities (e.g., Closed Loop automation and orchestration) enabled by end-to-end programmability of the end-to-end networks (up to the devices level). In this context, the 6G networks will deliver novel 6G features, as described in Table 1.

Table 1: 6G main features description.

6G Features	Description
DIGITAL INCLUSION	6G design should facilitate affordable coverage of sparsely populated areas. User interfaces for 6G should be easy to use and intuitive. Digital inequality in accessing 6G services should be prevented.
ENERGY EFFICIENCY	Energy efficiency of mobile networks has improved but the total energy consumption has increased due to increased data transport volume, which negatively impacts the environment, adds cost to operations and threatens scalability. Improvements in energy efficiency for 6G must exceed forecasted growth in traffic volume, and energy consumption figures must be comparable and available at all levels of the system for system-wide monitoring and optimization. Network features must support low energy consumption of end-user devices.

6G Features	Description
ENVIRONMENTAL IMPACT	Service providers are expected to report on their Environmental, Sustainability, and Governance (ESG) performance, including their impact and that of their supply chain on natural resources, pollution, and waste. The overall environmental impact of 6G should be minimized by monitoring and adapting RF emission levels, monitoring resources consumed during manufacturing, monitoring greenhouse gas emissions over the complete life cycle of equipment and terminals and monitoring the impact of real-estate assets. Common indicators should be used in 6G design to allow for comparison and to facilitate the elaboration of the environmental impact of 6G-based services.
NATIVE TRUSTWORTHINESS	Public telecommunication networks are critical national infrastructure and require native trustworthiness to protect against unintended and unauthorized access, ensure personal and sensitive data is protected, and ensure stable and predictable performance. 6G must support trustable networking environments, including guaranteed service and data availability, security, and privacy across multi-party infrastructures, even in untrusted environments. 6G design should provide means to ensure high levels of resilience, security, safety, reliability, and privacy.
AUTOMATION OF END-TO-END SERVICE DELIVERY	Network operators will need to offer a larger and more complex portfolio of services to provide their customers with the best-fitted services for their needs. Automation of service delivery is critical for network operators to efficiently manage large portfolios of services, optimize service delivery time, and reduce the risk of errors. To enable automation, 6G networks must support standardized monitoring, different configurable trade-offs between optimization objectives, and flexible and configurable network functions.
ARTIFICIAL INTELLIGENCE AND COMPUTE RELATED CAPABILITIES	The ability to provide AI as a service (AIaaS) is an important aspect of 6G systems, which should benefit AI applications thanks to their ability to support large AI models, data sharing capabilities, and large-scale distributed learning. 6G system will need to provide specific capabilities, including high uplink traffic capacity, privacy and transparent handling of sensitive data, and efficient and reliable results for AI inference. Support the management of model training, deployment, and storage, as well as the coordination of AI capable computing resources.
HYPER-SPEED AND ULTRA-LOW LATENCY	6G networks should offer extremely high data speeds and ultra-low latency, which will enable a range of applications such as real-time remote surgery, immersive virtual and augmented reality, and autonomous vehicles.

6G Features	Description
SPECTRUM	6G networks will require access to a wider range of the spectrum than previous generations of wireless networks. Development of new technologies and techniques to make more efficient use of existing spectra and to enable the use of new spectrum bands. UWB frequencies can also be used for 6G communication if the corresponding technical solutions are available.
MASSIVE CONNECTIVITY	6G networks must support massive connectivity to enable the Internet of Things (IoT) and other emerging technologies. Machine learning algorithms will be used to manage and optimize network resources to support large numbers of devices.
BACKHAUL SOLUTIONS	Various backhaul solutions are required for large cells, relay technology, and satellite technology. The advanced backhaul connections can be done via: • mmWave and THz frequencies, including novel antenna solutions. • Visible line and power line backhaul solutions. • Satellites and mega-cells
INTELLIGENT RESOURCE ALLOCATION	6G networks must support intelligent resource allocation to optimize network performance and improve user experience. Machine learning algorithms will be used to allocate network resources such as spectrum, power, and computational resources in real time.

The 6G network capabilities and features are extended from today's 5G existing available services and network features. These new capabilities include: 6G network resources scheduling, flexible and elastic self-services infrastructures, management through orchestration, fast adaptability to the services requirements, E2E control loop.

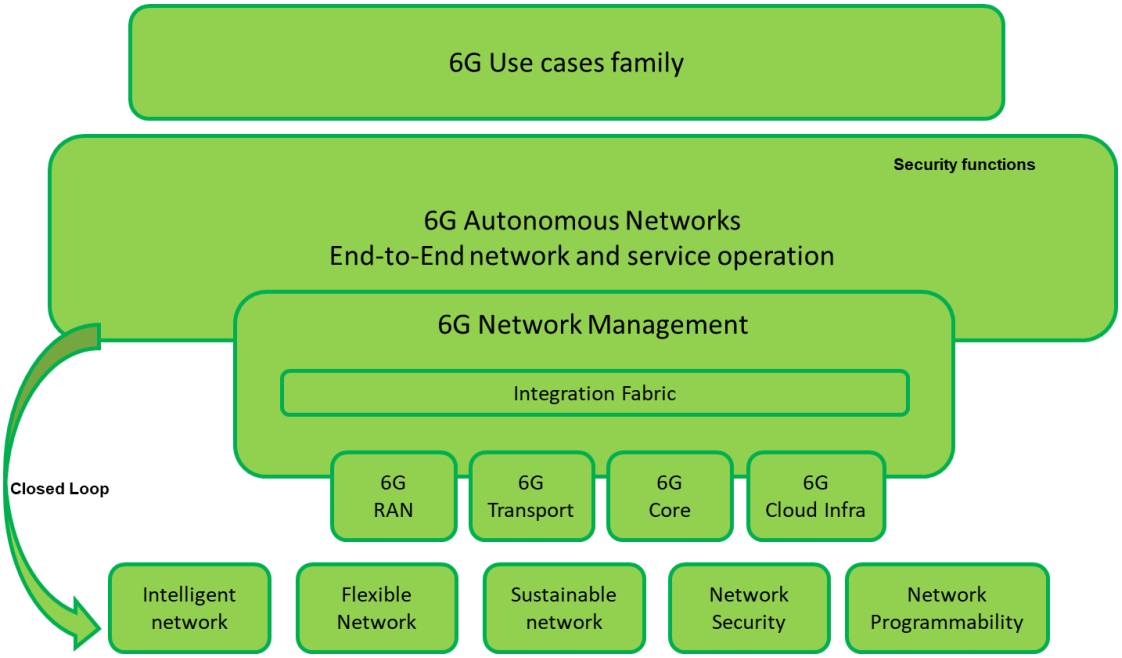

Figure 3: 6G networks requirements and features

As depicted in Figure 3, 6G supports the feasibility of new 6G Use cases family, mapping vertical's Use Case requirements to the available network functions and tools through a **Management and orchestration** layer, that – with the support of AI/ML and data driven techniques - takes care of services creation and provisioning, QoS and QoE fulfilment, performance and fault reporting, scaling and TMF Full Autonomous Network capabilities, etc. This layer can leverage on a pool of key 6G features that includes:

- **6G Intelligent network** is the autonomous and adaptable - with no human intervention - network instantiated and deployed based on cognitive and closed-loop functions, with support of AI algorithms. The intelligence in the network is required to be supported by the architectural design in term of components and services for the end-to-end 6G network implementation.
- **Flexible Network** represents the capability of the 6G network to adapt to different functionalities and technologies, increasing the availability and the reliability of the network.
- **The Sustainable 6G** network is the available and capable network to support the environmental, social and economic aspects, based on cost efficiency, sustainability and energy consumption, targeting the non-functional 6G KVIs for the key technological enabler's partners.
- **6G Network Security** is another important enabler for the future network's implementation, focusing on security and cyber security aspects, as resiliency against attacks, privacy, threat prevention and active protection mechanism through AI. The security is defined within the 6G architectural framework, and it is applied to all network components elements.
- **Network Programmability**, the vertical-oriented native interaction with the network through an API framework, 3rd party application onboarding, policy enforcements and network function monitoring. The 6G technology is the enabler for the exposure of network resources for several scenarios, as network deployment and connectivity level, resources management level and services and application's provisioning level.

3.3 Service-Based Architecture

The Service-Based Architecture represents the real epicenter of the 6Green architecture, in accordance with the Smart Network and Services (SNS) call: "*The work focus is on an architectural transformation, targeting energy-efficiency, using the flexibility that the Service-Based Architecture (SBA), introduced in Rel15 (TS23.501), is offering.*" But its relevance is not limited to the Work Programme, as the current SBA features make it very suitable to support the integration of the mechanisms and policies that are required to foster the ground-breaking innovations (e.g., *Edge Agility*, *Green Elasticity* and *Energy-aware Backpressure*) and to achieve the Project's targets. In particular, the 5G core being (micro-)service-based allows for the design and evolution of the individual NFs, as well as a smooth interaction among them through RESTful APIs and flexible and efficient network slicing. In addition, its centrality in the 5/6G ecosystem, with the potential to interact with both the infrastructure and the vertical application domains, makes it particularly fit to be a gathering point for the collected metrics, making decisions and delivering them to the other stakeholders to propagate the backpressure.

So, the design of the 6Green SBA (Figure 4) actually entails the development of the individual NFs, of the monitoring and profiling mechanisms, of the actuation policies at the network platform level, and of the management and exposure functionality to ensure that the actions are properly propagated across the whole ecosystem. In order to introduce all of these features and properly describe them, the preliminary design of the 6Green SBA, reported in the Project proposal, divides the NFs into "frameworks", i.e., groups of functions that work together by receiving data from the infrastructure/other frameworks, elaborating them and delivering the outcome to feed either other frameworks or the backpressure itself.

The frameworks composing the 6Green SBA are the *enhanced observability, monitoring, and analytics framework*, *the AI-driven decision and operations framework*, the *enhanced control policy framework*, the *integrated management subsystem* and the *management framework for edge-cloud resources*, as well as an *exposure layer* towards the verticals.

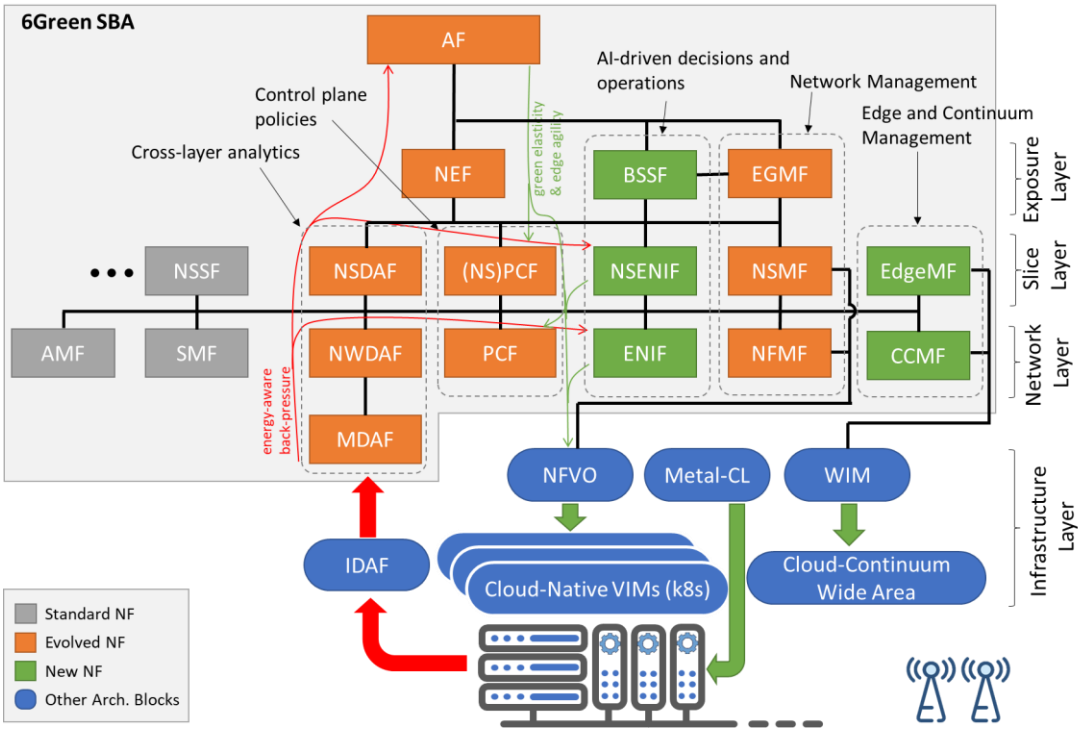

Figure 4: Preliminary design of the 6Green Service Based Architecture, as reported in the proposal.

In a nutshell, the *AI-driven decision and operations* framework consumes the backpressure data coming from the *observability, monitoring, and analytics* framework and the *management framework for edge-cloud resources*, and triggers edge agility and green elasticity operations actuated by the *control policy* framework and the *integrated management* subsystem. The presence of an exposure layer reinforces the bidirectionality of the ecosystem by propagating KPIs and requests for reconfigurations to/from all the involved stakeholders.

Before going into details on the individual frameworks and the NFs composing them, it is worth highlighting that, while the main roles and interactions will likely stay the same, the extension of existing functions and the design of brand-new ones will likely be revised during the project's lifetime. At the time of writing, it is envisaged that all the NFs will have an external state, in accordance with the serverless paradigm, to simplify their lifecycle management by allowing the NFVO to push and pop session data among the geo-distributed function instances. When useful, the hardware offloading of some of the NFs will be taken into consideration, this will likely be a viable option for AI-driven NFs as well as for the UPF.

The *observability, monitoring, and analytics* framework deals with the collection of metrics from the infrastructure and the other SBA NFs and with their synthesis into KPIs suitable to be delivered to the verticals. It is clear that the extension of the functions composing this framework, namely the Management Data Analytics Function (MDAF), the Network Data Analytics Function (NWDAF) and the Network Slice Data Analytics Function (NSDAF) regard the "green" side of the metrics collection and analysis, spanning from the collection of energy- and carbon-metrics coming from the infrastructure to their breakdown on a per-slice/vApp basis.

The MDAF has been defined in 3GPP TS 28.533 [28.533], to retrieve Operations, Administration and Maintenance (OAM) data and to perform analytics, as well as provide predictions, on a per-NF or per-slice

basis. Since the current definition already contemplates predicting resource usage according to the load level or resource utilization associated with the SBA NFs, its evolution in 6Green will extend, on the one hand, the plethora of collected data to the energy context, and on the other hand it will be fed by an Infrastructure Data Analytics Function (IDAF) designed to collect hardware-level metrics from the computing infrastructure of each VIM towards the 5/6G SBA. By combining metrics coming from both the infrastructure and the management, the MDAF will be able to estimate the current and the future carbon/energy footprint induced to the computing and offloading resources in the edge-cloud continuum, and even the availability and the use of green energy sources and hardware offloading engines.

While the implementation of the MDAF has been minimally explored so far in the literature [HCC22], the NWDAF on the other hand is collecting more interest [CMS22], [MCS22], likely due to its association with the ETSI ENI specification [ENI012]. The NWDAF is conceived to acquire monitored data at the edge and the cloud part of the infrastructure and to produce and analyse added-value KPIs, like forecasts of NF workloads, etc. Data collection is performed by subscription to events provided by network functions.

The evolution of the NWDAF, in line with what anticipated for the MDAF, will deal with the acquisition of control-plane data and their elaboration along the management ones coming from the MDAF into KPIs able to characterize the energy-efficiency level of the current SBA deployment and make forecasts accordingly. A prototype of the NWDAF NF has already been implemented by CNIT with a modular, microservice-based structure composed of a metric collector, a storage and an analytics module. The latter module will be particularly relevant to 6Green as it allows to easily plug the algorithms that will be designed during the project's lifetime.

The NSDAF has a similar role with respect to the NWDAF, but it is devoted to the elaboration of the NF-level KPIs provided by the NWDAF into figures ascribable to a single slice. This duality between functions operating with similar roles at NF-level and at slice-level is a *leitmotif* of all the frameworks, as will be punctuated in the following. For this function, along with the definition of energy-related KPIs, the evolution will cover the geographical scope of the edge-cloud resources devoted to host vertical applications.

It is evident that AI is the main enabler for this framework: the data collected by the MDAF and the NWDAF will be crucial for the training of the AI agents in charge of forecasting computing resources usage, traffic load and related induced footprint. But another aspect in which the role of AI is crucial is the decision-making process that is initiated by these forecasts. To this end, we have included an *AI-driven decision and operations* framework in the design of the 6Green SBA, which will be developed in compliance with the ETSI ENI specification but with the addition of multi-objective "green" policies aimed at supporting energy-aware operations. Such policies will be onboarded dynamically, to enable the flexible customization of the network and of the slices according to varying energy and carbon targets.

The usage of AI techniques for power control is not innovative, but the application is usually for the power control of sensors in wireless systems [ZYH+20] or to monitor the consumption of some aspects of the AI itself, such as training [MCD22]. However, the most commonly used AI/ML techniques, such as Deep Reinforcement Learning (DRL), will definitely provide the foundation of the application of these techniques to the collaborative control and management of infrastructures, networks and applications.

The framework is composed of two brand-new NFs, the ENI Function (ENIF) and the Network Slice ENI Function (NSENIF), that operate on the overall network and within a network slice, respectively. These functions will operate in a harmonized fashion by feeding each other with their results and ensuring the fulfilment of the potentially conflicting requirements of the vertical industries relying on the network platform and of the network itself. In their design, attention will be taken to make sure that their implementation will not turn out into a bottomless pit of consumption; fortunately, this topic is well-explored in the literature [GRR+19] and the available solutions will be taken into consideration for our work.

ENIF puts together the information coming from the observability, monitoring, and analytics framework, from the management framework for edge-cloud resources and from the network slices. Accordingly, it delivers requests for edge agility and green elasticity operations that involve the reconfiguration of the SBA NFs (e.g., horizontal/vertical scaling, migrations, etc.) and of infrastructural components, for example by enabling hardware accelerators for offloading purposes. The related control plane operations, such as the redirection of traffic towards the activated offloading engine, are also driven by ENIF.

NSENIF will play a role similar to the one of ENIF, but its scope will be the reconfiguration of an individual network slice and of the related applications. As such, it will be fed by the NFs involved in the analytics and edge management functions at slice level, such as the already mentioned NSDAF. In more details, the policies of a NSENIF instance will be devoted to making decisions on the NFs composing a slice, on the application components relying on it and on the hosting hardware in the edge-cloud continuum. The operations requested by the NSENIF may deal with the lifecycle of the NFs composing the slice (for instance, creation/termination of a UPF), as well as migration and related traffic steering. Of course, the design of the NSENIF will also ensure the support of multiple policies that can be dynamically plugged according to the current target.

The effective actuation of the AI-driven decisions is performed by three further frameworks operating at the control, management and edge-cloud resource level, respectively.

The **enhanced control policy framework** is composed of two functions, the Network Slice Policy Control Function (NSPCF) and the Policy Control Function (PCF), in charge of providing policy rules for control plane functions (such as subscription information, roaming and mobility management) according to the required QoS at slice and platform level, respectively. These functions will be evolved with the ability to handle the requests for control plane reconfigurations coming from the ENIF and the NSENIF. This mainly entails the development of the APIs between the two frameworks, for this reason it will be of vital importance to proceed with the development of all the SBA NFs in a parallel fashion as much as possible.

Regarding the management operations actuation, it is performed by the **integrated management** subsystem. This framework is heavily based on the 3GPP TS 28.533 [128.533], which defines the management of network functions and services through operations on their lifecycle (e.g., creation, update and deletion of NFs and NSs), performance, configuration and fault. The specification defines the interactions between “producer” and “consumer” management entities and contemplates the interaction with the NFV MANO and ETSI ZSM frameworks [ZSM002].

The extensions to the Network Slice Management Function (NSMF) and the Network Function Management Function (NFMF) will be performed in the direction of the Edge Agility and Green Elasticity innovations: thanks to the NFs design as cloud-native microservices, that facilitates the NFVO operations across the edge-cloud continuum, it will be possible to consolidate functions and slices where they can be run in a more efficient fashion (e.g., where renewables are available or their migration allows to turn off servers), or to off-load their workload on acceleration engines.

In addition to the management operations at platform and slice layer, this framework also includes, as intersection with the *exposure layer*, the Exposure Governance Management Function (EGMF). This function is crucial for the efficacy of the energy-aware backpressure because it provides the verticals (by interfacing with the AF) with the capability of intervening on the edge-cloud resources hosting their application components, including hardware offloading engines. The design of this function will stem from part of the APIs already developed between the Network Application Orchestrator (NAO) and the OSS (the remaining part will constitute the backbone of the Application Function, AF); the main differences will regard the structure of the code, which will pass from being a module of the OSS to a cloud-native microservice, with all the advantages given by the flexibility of this structure.

In addition to the actuation of control and management operations, the introduction of the 6Green solutions cannot neglect the role of the computing/networking/acceleration devices in the edge-cloud continuum, as they are crucial to achieve the actual energy-savings that the project is targeting. To this end, a management framework for edge-cloud resources, composed of two brand-new SBA NFs, has been conceived for providing verticals and network platform providers with the ability to intervene on the hardware devices hosting applications and NFs components.

The **management framework for edge-cloud resources** is foreseen to be composed of two NFs. The Cloud Continuum Management Function (CCMF) is in charge of providing the *AI-driven decision and operations* framework with the knowledge of the wide-area continuum topology. Such knowledge is obtained by interacting with the Wide-area Infrastructure Manager (WIM) and the component that maintains the topology of the edge-cloud continuum (including VIM resources). The combination of the topological information with the resource usage coming from the MDAF helps drive the ENIF/NSENIF decisions. The exposure of the related computing resources to the verticals is performed by the Edge-cloud Management Function (EdgeMF). This function, along with the EGMF, is the main responsible for the energy-aware backpressure, providing the verticals with the capability to actually intervene in the joint negotiation of the resources and so to actually be an active player in the green business/use models and Decarbonization Layer Agreements.

The 6Green SBA is completed with an *exposure layer* that allows interacting with the AF for the (re)negotiation of the network slices. The extensions in this case will be done to enable the cross-domain interactions with the involved stakeholders to support not only the slice-level edge agility and green elasticity operations, but also to enforce the business side of the ecosystem, by equipping the NFs with intent-based interfaces and meta-models allowing the stakeholders to cooperate within the green economy model and the joint decarbonization targets. Basically, the *exposure layer* is composed of the Network Exposure Function (NEF) and the Business Support System Function (BSSF), but its workflow is heavily supported by the already defined EGMF and by the AF.

The AF will be designed starting from the intent-based interface developed in the MATILDA [MAT17] and 5G-INDUCE [5GI21] projects by UBI and CNIT to enable the interactions between the verticals and the network platform to compose and manage slices. The extensions foreseen at this stage regard the interaction with the *exposure layer*, with specific APIs devoted to observing KPIs, renegotiating network slices and edge resources, including offloading engines.

The NEF is in charge of exposing the SBA to third parties by interacting with their AF in a bi-directional manner, which means that it takes care of abstraction and security of data to and from the AF. The evolution in 6Green will regard both the exposure to the vertical of the energy-aware backpressure KPIs coming from the NWDAF/NSDAF and the BSSF and the reception from the vertical of requests for edge agility and green elasticity operations. The fact that the usage of a CAPIF interface has been specified by 3GPP is very useful for 6Green because it perfectly complements the intent-based interface that will be established between the NAO and the AF. The south-bound interface of the latter, on the other hand, will be conceived from scratch following along the development of the other NFs, especially NWDAF and NSDAF.

Finally, the BSSF is a newly introduced function that will handle the business side of the network and slice reconfiguration, that is, the enforcement of the dynamic green business model between Telecom Operators and Vertical Stakeholders. We are aware of potential juxtapositions in the roles of the NEF and the BSSF, as they both involve operations on network slices and edge-cloud resources; at the time of writing, we have not completely decided whether the operations will be completely separated among the NFs or if the same operations triggered by different KPIs/actuators will be performed by both the NEF and the BSSF. This issue will be one of the first activities to be undertaken within WP3.

4 Use Cases and Scenarios

4.1 Introduction and Methodology

The purpose of this section is to define reference use case scenarios, along with measurement methods, metrics, and reference values. These elements are essential for evaluating the energy efficiency levels of the 6Green framework. We will analyse three distinct use cases in the following subsections:

- Use case 1 - Critical Operation Maintenance during Energy-constraint Disaster Scenarios
- Use case 2 - Energy-Efficient Augmented Reality Remote Assistance System
- Use case 3 - Zero-Carbon Clientless Virtual Enterprise Desktop as-a-Service

Each scenario is accompanied by a general description that provides context for the use case, including its underlying architecture, the associated business framework, and the relevant stakeholders involved.

Following the introduction, we will delve into the technology and functional enablers specific to each use case. The goal is to clearly identify the hardware and software functionalities that form the foundation of the use case and ensure its feasibility. We will consider both well-established technological features and novel ones developed specifically for this project.

To assess the performance of the 6Green solution within each use case scenario, we will identify a set of relevant Key Performance Indicators (KPIs). For each KPI, this document provides:

- A description, which provides an understanding of the parameter being measured.
- The objective, clarifying the purpose of optimizing the KPI within the specific context. In some cases, KPIs are grouped based on a common objective, facilitating their roles in the analysed use case. Examples of general objectives include network and service availability, critical services prioritization, network latency, carbon reduction, and more.
- The measurement methodology, specifying the adopted procedure and any underlying conditions or assumptions.
- The unit of measurement.
- A target value, representing the desired value that ensures proper functioning of the 6Green platform within that specific scenario.
- The expected impact on energy efficiency, indicating - when it is possible - how the KPI is anticipated to affect energy consumption within the 6Green framework.

4.2 UC1: Critical Operation Maintenance during Energy-Constraint Disaster Scenarios

4.2.1 Use Case 1 Description

User story: The term “critical infrastructure” relates to the infrastructure essential for the functioning of a society and economy, and would commonly refer to energy and utilities, information and communication sector, transportation, water supply, etc. Therefore, any interruptions of critical infrastructure operation, being either due to natural or man-made hazards, would have obvious consequences. Most of critical infrastructure operations are enabled by digital means and therefore tends to be designed and deployed by several principles already minimizing potential damage, e.g., redundant capabilities, distributed capabilities, uninterruptable power supplies, multiple alternative communication links, etc.

UC1 will assume distributed computing capabilities (e.g., edge continuum), including non-public 5GAdvanced/6G network as a main communication option supporting critical infrastructure operations. The use case will focus on maintaining operations in case several parts of the computing capabilities are unavailable (e.g., out of power, destroyed, sabotaged, etc.) or may become unavailable due to non-infinite power supply redundancy (e.g., electrical grid is offline, while local battery and solar powered backup is still available). Let assume a situation where central cloud location is off for some reasons, while only certain edge and far-edge nodes remain operational (see Figure 5), causing remaining capabilities to be unable to fulfill all requirements. In such a situation, certain functions need to be deprioritized, moved to other execution resources, or even stopped, with the final goal of minimizing consequences for the customers, i.e., trying to keep most crucial functions operational. In fact, such measures need to be employed as soon as there is a risk the situation would not normalize in the timeframe in which redundant capabilities are able to cover the outage of the affected resources.

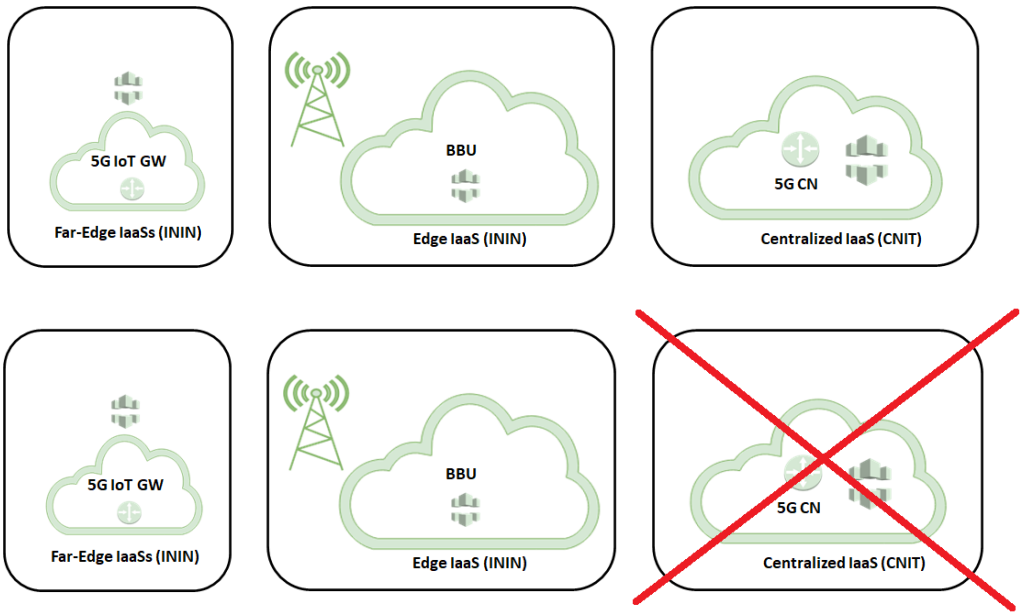

Figure 5: Network architecture during normal operation (top) vs.network architecture during disaster scenario operation (bottom).

Objectives: While researching, developing, and innovating on increasing energy efficiency of ICT technologies on one side, the UC1 will focus on evaluating possible consequences of the disaster events and on identifying feasible countermeasures aiming to preserve stability of the complete critical infrastructure system, thus reducing negative impact for the customers, i.e., society and economy, as the primary objective of the use case. By collecting multiple data in real time, as well as considering historical data, feeding them all into in AI/ML algorithms, the objective is also to evaluate current and future state of the system. Certain challenges while achieving use case' objectives have been already identified:

- constraint energy environment with weather/cloud dependent energy source (solar panels),
- limited battery capacity of portable power station,
- evaluating potential consequences of the disaster event,
- providing timely response of the system,
- executing proper actions during disaster scenario,
- restoring normal operation of all components when disaster scenario is over,
- complex trade-off policies between system usability and employed energy saving policies.

Architecture: The testbed (Figure 6) will utilize the CNIT platform as a central cloud. Then, the ININ's 5G IaaS platform (BBU components, RRH, antennas) will serve as an edge node. The ININ's industrial IoT gateway, connected to the 5G network (6G in the future), will be used as a far-edge node, i.e., providing IaaS capabilities for 3rd party applications (either virtualized or containerized). The edge node will be further equipped with an additional portable power station powered by solar panels. Next to this, drone mounted video cameras and fixed video cameras will provide real-time video surveillance as a security process supporting critical infrastructure operations, e.g., detecting potential physical attack and failures of the infrastructure, detecting other (non-intended) threats, etc. Application components (preferably cloud-native), crucial for the critical infrastructure support (e.g., 5G network components, AI/ML video analytics), will be distributed all around the testbed facilities (based on the function it serves, i.e., following 6Green Cloud Agility paradigm), though strongly synchronized to allow for proper redistribution during the disaster scenario. Following the 6Green Green Elasticity paradigm, actions in disaster scenario is expected to include reducing communication capabilities (e.g., reducing bandwidth, dropping non-prioritized users and traffic, powering off certain RRHs) and controlling states of processors on the nodes to adapt electrical power consumption and/or computational power for critical processes.

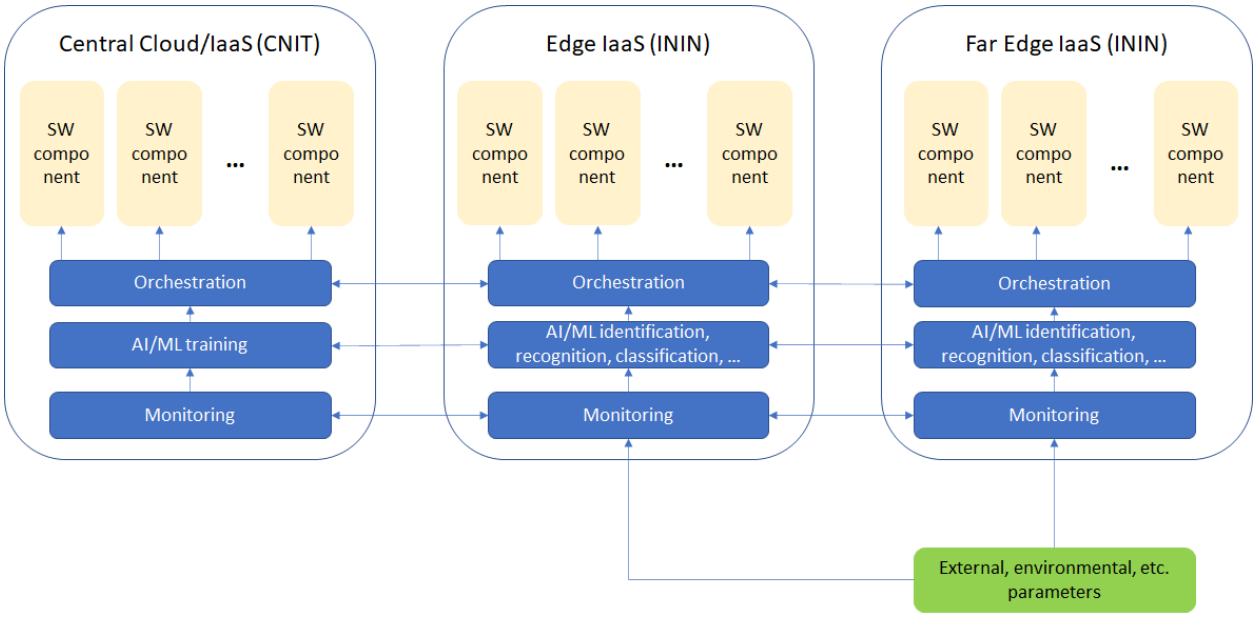

Figure 6: Operating in constrained energy environment – monitoring components collect multiple parameters feeding AI/ML algorithms to decide and push via orchestrator which SW components (those enabling 5G/6G operations and user services, i.e., critical infrastructure video surveillance) should run where and how, i.e., following the 6Green paradigms - Observability, Edge Agility, and Green Elasticity.

Business case: The technical solution addresses the fact that, as it appears clearly in recent times, there is no guarantee that anything, yet unimaginable, can happen, e.g., war, cyber-attacks, terrorist attacks, natural hazards, etc., and in that scope the use case provides flexible operation of the ICT system during disaster scenario. Although there are various solutions already available in the market (e.g., redundant capabilities, distributed capabilities, uninterruptable power supplies, multiple alternative communication links, etc.), the UC1 solution provides an alternative view and immediate remedy for the issue. Additionally, the solution includes local solar powered power station which serves as a backup in the case of emergency, as well as local source of "green" energy. In combination with optimizing energy consumption (especially during disaster scenario operation), this targets environmentally friendliness and consciousness as added values of the solution.

Relevant stakeholders: The use case addresses critical infrastructure operators, such as telco operators, and aims to propose possible solutions for enhancing their operations using the 6Green approach combined with the latest technology. Likewise, critical infrastructure users can explore new use cases related to their particular needs, suggesting additional improvements in the critical infrastructure and services operation under specific energy-constraint conditions. Service providers and technology vendors can also seek further enhancements for the proposed solutions, and work towards transforming the concept into tangible, real products. Additionally, standardization and legislative bodies should evaluate the societal benefits of these solutions to harness the improvements for the entire society.

4.2.2 Use Case 1 Technology/Functional Enablers

Central cloud and 5G/6G CN

In this section, we give an overview of the Core Network (CN) technological key enablers that may be relevant to this use case.

In situations where critical operations must be performed in disaster scenarios, the capabilities of a CN acquire a crucial role for maintaining network operations at the edge. Assuming that most of the core functionalities could be deployed in the cloud, an architectural flexibility to offload CN functionalities to the edge and to ensure full connectivity in case of a central cloud outage, is a must.

In recent years, research and standardization plans of 5G/6G system have focused on solution approaches for many verticals and have provided, considering the CN side, a series of functionalities and characteristics that allow multiple architectural and deployment models capable of adapting to different needs and environments.

A 5G/6G mobile core network is introduced as an open and modular core network services platform described by the Service-Based Architecture, as seen in Figure 7, which provides a framework of interconnected Network Functions (NFs) accessing and interacting to each other adopting a Service-Based Interface (SBI). This SBA paradigm promotes cloud-native architectures, supporting different NFs or software applications running in the cloud infrastructure as Virtualized Network Functions (VNFs) or Containerized Network Functions (CNFs).

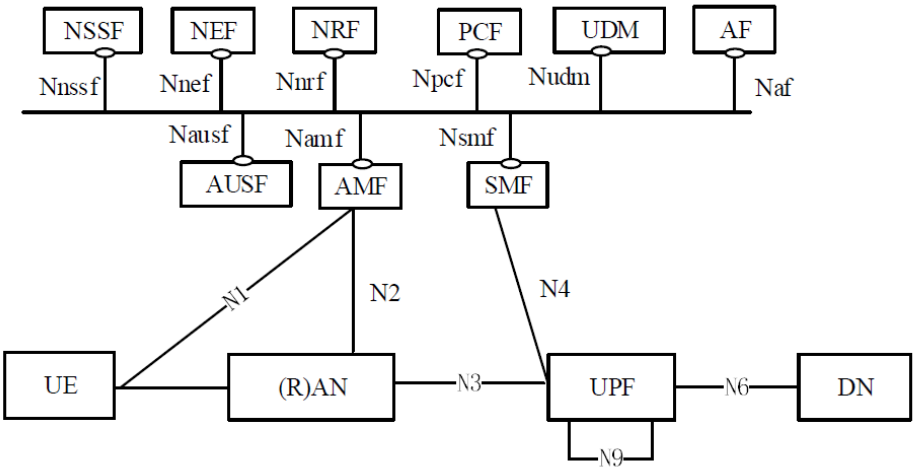

Figure 7: 3GPP 5G System Architecture.

Along this direction, ***virtualization* of NFs** is a peculiar paradigm for architecture flexibility and customization, since different VMs or containers can be moved to other Network Function Virtualization Infrastructures (NFVI) hosts. The 5G system architecture enables services deployments based on different virtualization techniques (e.g., IaaS/CaaS) and leverages on services-based interactions, User Plane and Control Plane separation, scalability, dynamic and flexible modularized deployments, distributed or centralized solution for XaaS model. The SBA enables NFs to interact with other NFs in different communication models, direct or indirect, based on the 5G intelligent service traffic control between NF consumers and producers, using APIs, as described in Table 2.

Table 2: NF/NF consumer-producer communication model.

Communication type	Services discovery	Communication model
Direct communication	Direct NF routing	A
	Discovery using NRF services, no SCP; direct routing	B
Indirect Communication	Discovery using NRF services, specific SCP delegation, SCP routing	C
	Discovery and selection delegated to an SCP, SCP routing	D

The envisioned reference model is defined as a service-based representation of the cloud-native core network implementation, containerized NF core approach, NFs accessible to any other authorized NF through SBIs, as the services are exposed and described by using APIs specifications. The cloud native application approach decomposes the software into smaller and manageable pieces for the microservices architectures. As the 5G/6G Core NFs are enabled with APIs, on demand different CN functions can be deployed, publishing the existing services in the NRFs which are also providing services available in the 5G/6G network, based on the 3GPP Restful APIs over SBIs [29.501]. The 5G/6G Core currently implements the IT network principles, making the implementation effective and implementing novel life-cycle management, allowing the Management transformation path to the Zero-Touch-Operations, Network Automation based on AI/ML innovative frameworks, supported by cloud-native and edge computing infrastructures enabled with automated software development and flows for design, testing, integration, delivery, the CI/CD (Continuous Integration / Continuous Delivery) and DevOps approach for 5G/6G telco infrastructures.

In this context, and considering this specific UC, a **flexible virtualized 5GC deployment** model, with proper NFs arrangement and re-location approaches, can be orchestrated according to specific energy and service needs.

For example, a simple way to preserve connectivity of the complete critical infrastructure system is to instantiate a **back-up core network at the edge**, so that when the central cloud becomes unavailable, the corresponding redundant NFs' functionalities are conveniently reconfigured and energy-optimized for maintaining critical communication operations. Furthermore, it is possible to **prioritize the traffic** only for some critical services (such as those used by first-responders or military teams) making effective a preservation of energy availability only for selected users. Furthermore, an **energy-aware orchestrator** can operate on top of the core by **enabling/disabling some NFs** or related services when not needed, thus leveraging **management 5GC Open APIs**.

Portable 5G/6G system

The portable 5G/6G system (Figure 8) is a compact portable mobile node for field-based network testing and services verification, as well as for creating experimental sites in general. It was developed by the support of Horizon EU-funded program, and it has been used in several 5G-PPP projects. The system comprises of an SDR- and CPRI-based radio and mobile core system (4G and 5G for now) deployed on the hardware components installed in a box-size suitcase. It provides flexible configuration options powered by NFV, services, test, and validation toolkit, and is also ready to be connected to cloud backend infrastructure to provide additional capabilities.

Figure 8: Portable 5G/6G system – complete set-up including radio part components (left), and edge IaaS in compact form (right).

Portable power station

The portable power station (Figure 9) is a combination of batteries, packed in a compact portable box, and solar panels which produce zero emission electric power for charging batteries and supplying portable 5G/6G system when required. According to the use case scenario, the portable power station is aimed at providing energy during the conditions of energy shortage, i.e., energy-constraint disaster scenario. Operation of the system during the energy-constraint disaster scenario is impacted by limiting certain services/functions capabilities which will depend on the energy that the portable power station is able to provide. However, the latter necessarily depends on power provided by solar panels (daytime vs. nighttime) and remaining battery charge.

Figure 9: Portable power station (symbolic photo) consisting of batteries and solar panels.

Far-edge IaaS

The far-edge IaaS (Figure 10) is a powerful industrial grade device/computer with 5G/6G connectivity. Its main purpose is to serve as a 5G/6G gateway for non-5G/6G devices which, when connected to it (e.g., by USB, ethernet, WiFi, etc.), become accessible via 5G/6G network. These devices would normally be various IoT devices and sensors for industrial use, including usage in critical infrastructures. Far-edge IaaS processing and storage capabilities enable virtualization and running containerized software components providing, e.g., added value to raw data acquired by IoT sensors. Such architecture further enables orchestration of software components, thus also enabling steering services/functions according to requirements influenced by energy-constraint disaster scenario within particular use case.

Figure 10: Far-edge IaaS device.

Edge agility and 6Green Elasticity

Edge agility and elasticity are considered crucial characteristics for the lifecycle management of distributed application components for disaster management scenarios. Deployment, scaling and live migration of application components across resources in the computing continuum is required to support strict QoS requirements, especially in the far edge part of the infrastructure. Autoscaling mechanisms can be very helpful to achieve the assurance with SLAs in case of bursty workloads at the edge part of the infrastructure, considering the intensiveness of the consumption of resources in case of video streaming and AI/ML video analytics functions. Agility in terms of decision making of the various orchestration components is also important for the management of both network and cloud resources, since decision making in disaster management operations is time critical. Such agility will be also introduced with the objective to achieve high energy efficiency in the deployment and runtime of distributed applications and network services by examining the trade-off between performance and energy consumption. Scaling to zero and serverless mechanisms will be considered for the edge/far edge part of the infrastructure to support –on demand– strict QoS requirements.

4.2.3 Use Case 1 Performance Metrics

Network Availability

During the disaster scenario, energy constraints conditions are expected to happen. To provide required network availability for critical services, the solution proposed within the use case includes battery and solar panels supply. The same solution will provide availability of critical services during the energy constraints conditions. Since the energy provided by batteries and solar panels is limited, in Table 3 we propose a KPI suggesting the required network availability during energy-constraint disaster scenario.

Table 3: Network Availability KPI.

Network Availability	KPI_ UC1_1
Description	Percentage of successfully received data packages over the total packages transmitted via the connection path.
Objective (in scope of the Use Case context)	Critical operations rely on availability of the network at every time, i.e., also during the energy-constraint disaster scenario. The KPI will help design and optimise the network to meet the availability requirements.
Measurement methodology	The network availability is an estimated rate of successfully received packages vs. number of packages sent during the observed time window when transmitting packages (e.g., ICMP) from the point A in the network to the point B (and vice-versa) in the network. Point A would be a device representing service consumer (e.g., client), while point B would be a device providing the service (e.g., server).
Target	Preserving the same availability rate (e.g., 99,99 %) during both the energy non-constraint and the energy constraint scenario.
Expected impact on energy efficiency	Meeting target KPI value helps meet target values of certain low level KPIs (e.g., CPU, GPU, memory usage, remote storage access, etc.).

Critical services prioritization: Energy-constraint conditions detection time, Time to disable non-critical services, Time to disable non-critical UEs connected

During the disaster scenario, energy constraints conditions are expected to happen. To keep critical services run, certain tasks related to energy consumption optimization needs to be executed (e.g., this may also require shutting down certain non-critical services). Since the latter requires quick response of the system, or at least certain system components, we propose KPIs addressing and evaluating the time interval required for energy constraint conditions detection (Table 4) and time intervals to disable non-critical services and non-critical UEs that might be connected to the network at the time being (Table 5 and Table 6).

Table 4: Energy-constraint conditions detection time KPI.

Energy-constraint conditions detection time	KPI_ UC1_2
Description	Time interval required for the system to detect an energy-constraint conditions has occurred.
Objective (in scope of the Use Case context)	To enable service prioritization during the occurrence of disaster scenario, the system needs to detect energy constraint conditions has occurred, i.e., available energy resources does not guarantee unlimited operation time which could be, for example, due to interrupted energy supply through electrical grid.

Energy-constraint conditions detection time	KPI_ UC1_2
Measurement methodology	Test environment should provide means to simulate energy-constraint conditions. On start of the energy-constraint conditions simulation (e.g., disconnect power supply from electrical grid), also start measuring the time interval which lasts until the system successfully detects energy supply state change. Definition of the "energy supply state change successfully detected" will be provided during system design.
Target	30 seconds
Expected impact on energy efficiency	Prolongs availability of energy from batteries and alternative power sources thus adding to overall system availability and energy efficiency.

Table 5: Time to disable non-critical services KPI.

Time to disable non-critical services	KPI_ UC1_3
Description	Time interval required for the system to disable non-critical services.
Objective (in scope of the Use Case context)	To enable service prioritization during the occurrence of disaster scenario, it would be required in certain situation to disable non-critical services to keep critical services operating. As mentioned before, available energy resources may not guarantee unlimited operation time, while by disabling certain services, it might be possible to significantly prolong the operation time for critical services.
Measurement methodology	The pre-requisite for evaluating this KPI is successfully detected occurrence of energy-constraint conditions (KPI_UC1_2). As soon as the energy-constraint conditions are detected, it is possible to start measuring the time interval needed for the system to successfully disable non-critical services.
Target	30 s (for all non-critical services are disabled)
Expected impact on energy efficiency	Prolongs availability of energy from batteries and alternative power sources thus adding to overall system availability and energy efficiency.

Table 6: Time to disable non-critical UEs connected KPI.

Time to disable non-critical UEs connected	KPI_ UC1_4
Description	Time interval required for the system to disable non-critical UEs connected to the system/network.
Objective (in scope of the Use Case context)	To enable service prioritization during the occurrence of disaster scenario, in certain situation it could be also required to disable non-critical UEs still connected to the network. This would lower the energy consumption of the system and thus add extra time for critical services to keep operating.
Measurement methodology	The pre-requisite for evaluating this KPI is successfully detected occurrence of energy-constraint conditions (KPI_UC1_2). As soon as the energy-constraint conditions are detected, it is possible to start measuring the time interval needed for the system to successfully disable non-critical UEs.
Target	30 s (for all non-critical services are disabled)
Expected impact on energy efficiency	Prolongs availability of energy from batteries and alternative power sources thus adding to overall system availability and energy efficiency.

Network Latency - general description and introduction

In this paragraph we introduce an extensive and general description of the latency concept, the impact due to the various network segments and the measurement methodology. Subsequently there will be a specific description in the various UCs with reference to this paragraph for the aspects of general validity.

The latency of a communications network is defined as the time needed to transport information from a sender to a receiver. One of the most used measures of latency is the so-called Round-Trip-Time (RTT), which is defined as the time taken for a packet of information to travel from the sender to the receiver and back again (usually expressed in ms). Although we speak of latency as a finite term, it is actually an accumulation of delays which occur in different segments of a network. In other words, the RTT is the total time it takes for a packet of data to travel from the sender to the receiver, across multiple hops, plus the total length of time it takes for receiver to send an acknowledgment back to the sender, through multiple hops. In case of end-to-end latency, we must also consider the time for the application to process the request and this value depends on the application type.

We can also define a Client-to-Server or a Server-to-Client latency.

The main latency components in the Client-to-Server case are:

- Network latency on radio mobile network
- Network latency on core mobile network
- Network/Virtualization latency on Edge Computing
- Application process latency on Edge Computing

Similarly, the latency components in the Server-to-Client case are:

- Network latency on core mobile network
- Network latency on radio mobile network
- Network latency on Mobile Device/Desktop
- Application process latency Mobile Device/Desktop

As we are working in Edge Computing scenario, Edge Computing solutions are usually co-located in PGW/UPF sites and, therefore, the latency contribution due to transport is near 1 ms (one single switch/router hop). When considering other kinds of scenario, an extra-latency due to transport should be taken into account: it can be approximated as 5 µs per Km, and 1 ms for each switch/router hop.

We assume that the application is located at the Edge Computing solution that is collocated at the PGW in the case of 4G/5G NSA networks or at the UPF in the case of SA networks.

The latency contributions in Client-to-Server scenario of different network components are represented in Figure 11.

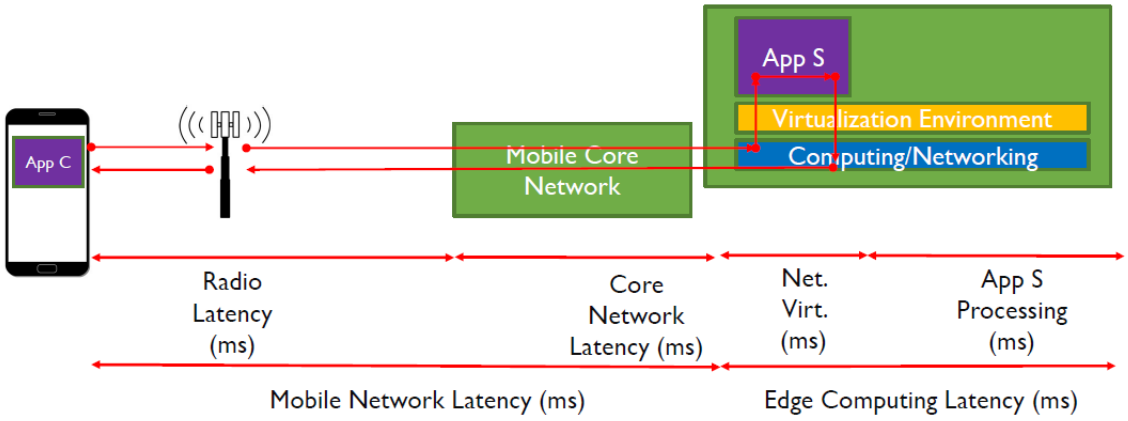

Figure 11: Latency Network Components contribution.

The latency measurement methodology proposed requires the introduction of 2 probes capable of tracing a specific protocol, linking requests with responses of a specific transaction and tracing timestamps. The probes (e.g., TCPDUMP) must be inserted on the mobile device side and Edge Computing side.

Considering a specific transaction (es. TCP request – TCP ack) the Time Stamps that will be measured are:

- T1 – First message (Request) sent from application on Client side
- T2 – Message arrived on Edge Computing Platform
- T3 – Message (Answer) out from Edge Computing
- T4 – Message (Answer) received from application on Client side

With these time stamps, the following latency components can be computed:

- RTT network = (T4 – T1) – (T3 – T2)
- RTT on Edge Computing = T3 – T2

In Figure 12 the schematic of the proposed Latency Measurement methodology is represented.

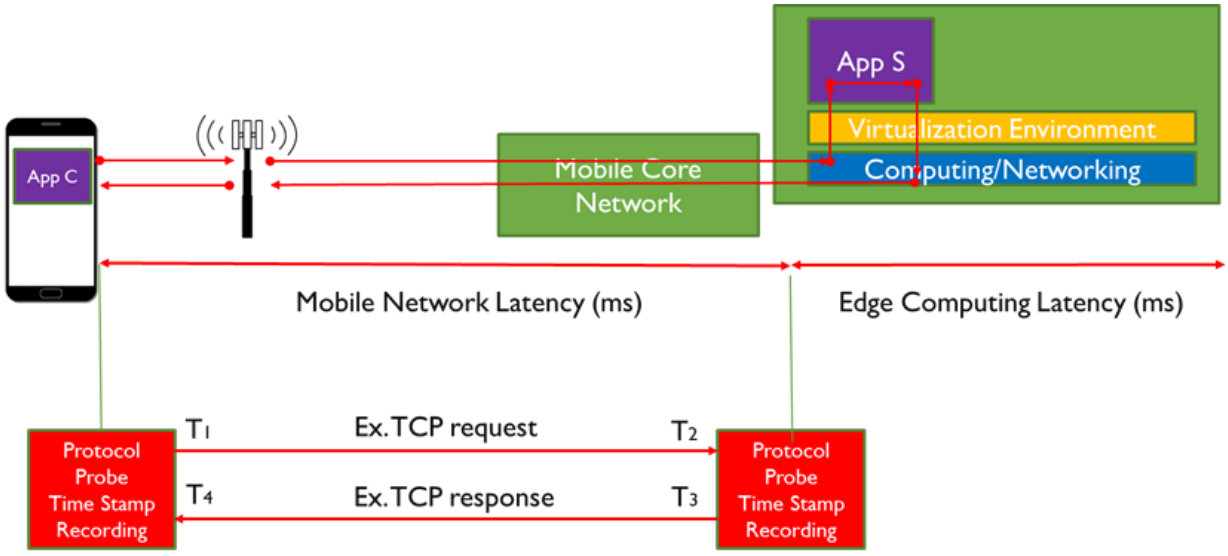

Figure 12: The schematic of Latency Measurement methodology.

In Table 7 typical values of Latency are reported in 5G NSA and 4G networks. The major contribution to Mobile Network Latency is due to Radio access and depends on Radio conditions. Contribution to latency of Core Network (NSA) is normally around 2-3 ms. Clearly the latency depends on different aspects, in particular:

- Radio condition where the Mobile Equipment is located
- Speed of Mobile Equipment
- Coverage
- Power management feature of radio side (e.g., DRX), since normally Power consumption optimization could affect latency.

Table 7: Typical Latency value in a 4G and NSA 5G.

	Mobile Network RTT [ms] (Radio + Core Networking)	Edge Computing Networking [ms]
5G NSA	Min ULL = 5 Min = 10 Max = 20	1-2
4G	Min = 16 Max = 40	

As for the latency requirements, these depend on the applications: typically, we start from around 20 ms for applications such as gaming, applications for self-driving cars, VR applications with remote control (e.g., drones, robotics). For more general VR applications, it's usually around 50 ms. In the case of applications such as VoIP and generic video, values of around 100 ms can also be reached. Regarding latency and energy consumption in general, especially on mobile networks and specifically on radio access, normally all technologies that are aimed at minimizing latency lead to an increase in energy consumption, especially in cases of low latency values.

Table 8 provides a breakdown of the latency KPI for the specific Disaster Recovery situation UC.

Table 8: Network latency KPI for UC1.

Network Latency	KPI_ UC1_5
Description	Server-to-Client network latency (depending on the specific app used)
Objective (in scope of the Use Case context)	In a disaster recovery situation, the latency for a 5G/6G mobile network may be impacted in several ways. Firstly, there may be a high volume of traffic on the network due to increased demand for communication and information sharing. This can lead to congestion and network slowdowns, resulting in higher latency. Secondly, damage to network infrastructure or power outages may cause disruptions to network connectivity, leading to longer latency times. Thirdly, the prioritization of emergency communications may take precedence over other types of traffic, leading to longer latencies for non-emergency traffic. Finally, the availability and location of network resources such as base stations, antennas, and backhaul links may be affected, leading to increased latency as traffic is rerouted to alternative resources. Overall, it is important to plan for disaster recovery situations and implement strategies to minimize the impact on network latency.
Measurement methodology	Latency (in ms) is measured from the client to the central deployment and back ("considerations on the processing time"). Probe based metric collection – client-server-approach. See par. 4.2.3 for further insights.
Target	20 ms (considering all non-critical services are disabled) This value strongly depends on the type of critical services under analysis.
Expected impact on energy efficiency	In numerous scenarios, optimizing latency often requires sacrificing power-saving mechanisms within the network, particularly for radio access. For applications that are less sensitive to latency, it is possible to loosen the latency requirements and achieve greater power savings. Conversely, for highly latency-sensitive applications, it may be worthwhile to consider implementing hardware acceleration systems for packet switching to ensure low latency while still obtaining energy savings.

Green energy maximization: Green energy utilization rate

During normal operation of the network (non-disaster scenario), the total emissions of a service can be minimized by means of energy-aware placement of workloads on resources that are powered by local low-emission power sources, as opposed to other sites where the energy sources available may be more contaminating. To measure this, it can be useful to define a KPI (see Table 9) that monitors the proportion of energy that a service consumes, which is locally produced with zero-emission.

Table 9: Green energy utilization rate KPI.

Green energy utilization rate	KPI_ UC1_6
Description	Percentage of the energy consumed by the end-to-end service that is produced locally by zero-emission energy sources.
Objective (in scope of the Use Case context)	The objective of this KPI is to monitor the proportion of locally produced zero-emission energy that is consumed by the end-to-end service during normal operation with the aim to compare the scenarios when the use of this energy sources is maximized versus the standard scenario when these resources are only utilized to maximize performance.
Measurement methodology	To measure this KPI, two metrics are considered: the consumption of the different parts of the service as a proportion of the total consumption on a per site basis using the instrumentation tools available (such as metered power distribution units (PDUs) and/or tools like Scaphandre), and the amount of available locally produced zero-emission energy.
Target	The target of this KPI is to observe a significant increment on the proportion of locally produced zero-emission energy used by the service. The amount of this increase will be dependent on the site´s energy production capacity and the weather conditions.
Expected impact on energy efficiency	It is expected that the carbon footprint of the E2E services is reduced significantly.

4.3 UC2: Energy-Efficient Augmented Reality Remote Assistance System

4.3.1 Use Case 2 Description

User Story: In 2017, 27% of European CO_2 emissions came from the transport sector. Business trips account for a large share of the carbon footprint in many companies acting in any vertical market sector. By replacing travel of expert technicians for advanced and augmented remote collaboration and inspection tools, companies can reduce travel costs and related CO_2 emissions and time spent.

oculavis SHARE is a Remote Visual Assistance application that helps machine or equipment manufacturers and manufacturing companies to optimize their remote services to clients. Experts do not need to travel

anymore to solve simple tasks or identify spare parts. All this can be guided and documented remotely by the expert supporting a technician on-site. The technicians on-site use smart glasses or smartphones/tablets to get connected. Reducing CO_2 emissions and machine downtimes by avoiding unnecessary travels are core goals of remote assistance applications. Moreover, companies will lower operational costs by improving resource use and reduce risks by increased access to competent personnel.

Objectives: the UC2 will show how state-of-the-art cloud-native technologies with new solutions can boost the overall ecosystem flexibility, scalability, and sustainability levels. On the one hand the UC2 will focus on possible ways to give use-case centered input for the 6Green SBA from the vApp oculavis SHARE (e.g., user A needs to connect to a server in Germany) to improve the Quality of Experience (QoE) for the end user, while on the other hand it will properly handle changes in the given Quality of Service (QoS) due to energy-efficient algorithms (e.g. change server location to run on renewable energy, shut down server to improve resource usage). The following objectives have been derived for UC2:

- Improve oculavis SHARE network architecture to get closer to their carbon neutrality by
  - Lowering resource use adaptive to its usage
  - Distributing server loads to reduce minimal resource use
  - Handling energy-efficient vApp deployment locations
- Lower operational expenditure by improving resource use of oculavis SHARE network architecture
- Adaptive handling of available bandwidth and/or desired bandwidth changes
- Handling of complex trade-off policies between QoS/QoE and energy saving policies
- Creation of new business models (e.g., eco-friendly Service Level Agreements)

Architecture: Remote Visual Assistance applications like oculavis SHARE deliver the necessary functionalities like an Augmented Reality videocall with annotation and documentation features. For delivering the software services, different server applications need to be provided such as database and web servers, STUN/TURN connection servers and media servers for establishing high quality video and audio connections all around the world. Figure 13 shows a simplified oculavis SHARE architecture and envisioned extensions for the 6Green project.

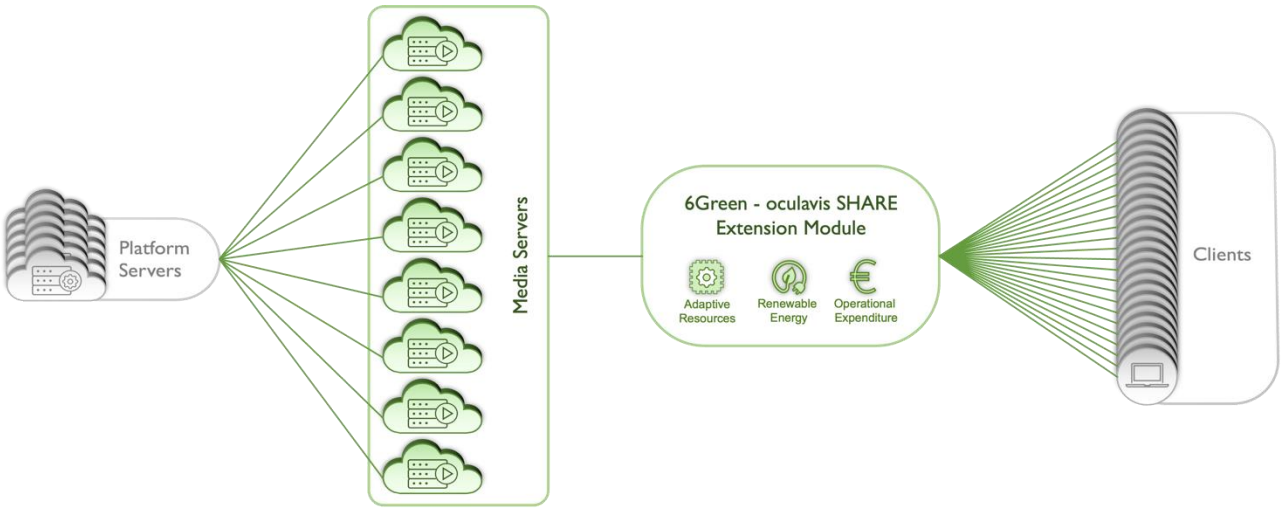

Figure 13: oculavis SHARE network architecture and envisioned extensions for 6Green SBA.

oculavis SHARE is a 'brownfield' software that requires certain basics and therefore the central architecture of the system in general cannot be changed due to business reasons. For instance, the database servers will

still be available in a central cloud environment and only media servers might be hosted for high quality and resource-efficient operations.

Business case: This vApp will be extended to fully taking advantage of energy-efficiency opportunities of the 6Green SBA and its functional and technological enablers. While the intention of the vApp itself is to reduce travel costs and CO2 emissions by applying remote visual assistance, server costs and produced CO_2 emissions are often neglected. Based on the project results and achievements, new business models including new eco-friendly service level agreements can be created and offered to customers.

Relevant stakeholders: The use case involves various stakeholders, including verticals, telco operators, infrastructure, and cloud providers. Vertical companies in the machine and plant industry are the end users and will benefit from the 6Green enhancements. These enhancements enable them to perform remote audits, inspections, and trouble shootings with higher QoS while contributing to energy-efficient solutions. Telco operators, infrastructure and cloud providers can explore new use cases and identify their specific needs. By doing so, they can suggest additional improvements to realize an energy-efficient augmented reality remote assistance system.

4.3.2 Use Case 2 Technology/Functional Enablers

5G CN Technology Enabler

The Remote Visualization Assistance application described in UC2 is a type of vertical that is gaining great popularity in the industrial and consumer world. A virtual/extended reality (VR/XR) application is characterized by multiple concurrent data streams, with different QoS requirements.

The impact on the market and the complexity of QoS requirements imposes the need for more advanced QoS management in 5G/6G networks, also from the point of view of the core network. Compared to 4G, 5G technology (on both RAN and CN sides) has also been designed to enhance support for QoS and QoE, adding new features and functionalities that characterize part of the next step called 5G-advanced, as illustrated in [LBK+23].

A CN capable of managing QoS in an advanced way offers a whole series of characteristics to adapt traffic flows to different service requirements. In the 6Green context, and specially in this UC, a 5G CN solution may be able to provide dynamic QoS management, essential for obtaining high flexibility in the use of available network resources. In simple words, a specific application (such as Remote Visual Assistance) can interact with the CN requesting a specific traffic priority or QoS to users belonging to the service itself, guaranteeing a secure and reliable service. This request is propagated to the rest of the 5G network elements, to ensure the assigned service along the entire E2E chain.

The dynamic management of QoS and resources allocation on the CN side can be demonstrated by leveraging the 6Green SBA architecture: SBA can dynamically require dedicated QoS levels according to different deployed services and available resources, exploiting open APIs exposed by the Policy Control Function (PCF) [23.501, 23.502, 23.503] or Network Exposure Function (NEF) [23.501, 23.502].

Cloud-native star topology media server architecture

The Augmented Reality Remote Assistance System oculavis SHARE is based on a star topology media server architecture. It is broken down into smaller services that are individually deployable and are interconnected in a cloud computing environment (Figure 14), which can be exploited by the 6Green project and its service-based architecture.

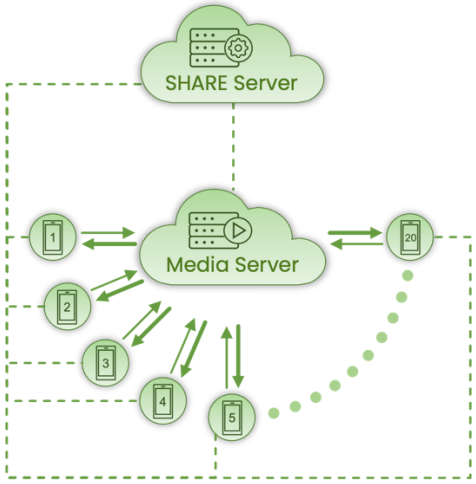

Figure 14: Cloud-native star topology media server architecture of oculavis SHARE.

Potential further developments to allow dynamic setup and (re-)configurations of the services will be achieved within the project to showcase the full advantages of the 6Green SBA.

Adaptive video stream layer handling

The native mobile applications of oculavis SHARE allow to manually configure the video stream layer (Figure 15) by setting video encoding parameters for high quality real-time communication beyond market standards of live video conferencing (e.g., max 1.5 Mbit/s for transmitting a maximum resolution of 720p in a multi-party conferencing scenario).

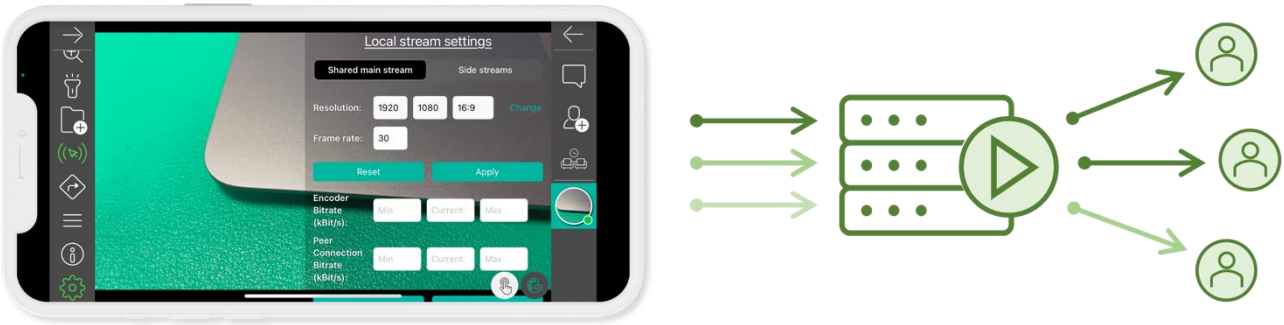

Figure 15: Adaptive video stream layer handling of native iOS oculavis SHARE application.

Based on industrial specific QoS needs for remote support use cases, such as remote inspections and audits of machinery and equipment, the video stream layer can be set up to 4K resolutions and achieve frame rates up to 60 fps given that the network provides the needed bandwidth.

Edge agility and 6Green Elasticity

Analog to Use Case 1, the edge agility and elasticity are also considered crucial characteristics for the lifecycle management of distributed application components for the augmented reality remote assistance system. Deployment, scaling and live migration of application components across resources in the computing continuum is required to support QoS requirements.

Autoscaling and migration mechanisms can be very helpful to achieve the assurance with SLAs in case of high usage and changing conditions of the provided green energy, considering the intensiveness of the consumption of resources in case of 4K video streaming.

These kind of agility and elasticity use case requirements will be introduced with the objective to achieve high energy efficiency in the deployment and runtime of distributed applications and network services by examining the trade-off between performance and energy consumption.

4.3.3 Use Case 2 Performance Metrics

Energy-efficient media server setup: Carbon reduction and Zero-Carbon Service, Operational Expenditure for media server cluster, Adaptive bandwidth

The usage of augmented reality in remote assistance systems can fluctuate a lot from multiple multi-party video calls to no usage at all during the night. Due to existing SLAs, the system availability needs to be guaranteed. To address and implement an energy-efficient augmented reality remote assistance system, server components need to be reconfigured (e.g., reserved vCPU resources of media server), unused server components need to be shut down or new server instances deployed and configured (e.g., additional media servers due to high load or making use of cloud infrastructure powered by renewable energy). To realize this, the system needs quick response times to keep the existing QoS. The following KPIs are proposed to evaluate the energy-efficient media server setup which is aimed in the 6Green project.

Table 10: Carbon reduction, Zero-Carbon Service Agreement KPI for UC2.

Carbon reduction, Zero-Carbon Service Agreement	KPI_ UC2_1
Description	While the intention of the vApp itself is to reduce travel costs and CO2 emissions by applying remote visual assistance instead of on-site visits, server costs and its produced CO2 emissions are often neglected. The KPI describes the carbon reduction achieved by 6Green energy-efficient media server deployments.
Objective (in scope of the Use Case context)	Improve oculavis SHARE network architecture to get closer to their carbon neutrality, e.g., by lowering resource use according to its usage, distributing server loads to reduce minimal resource use, handling energy-efficient vApp deployment locations and shutting down unused server under low loads.
Measurement methodology	Comparing the power (W) -and related CO2 emission- consumed by state-of-the-art oculavis SHARE network architecture running in standard cloud cluster compared to improved oculavis SHARE network architecture running in 6Green SBA.
Target	Decreased by 30%
Expected impact on energy efficiency	Reduced energy consumption due to less used server components and lower usage of CPU and RAM.

Table 11: Operational expenditure of media server cluster for UC2.

Operational Expenditure for media server cluster	KPI_ UC2_2
Description	The costs to run the media server cluster needed to apply an energy-efficient augmented reality remote assistance system.
Objective (in scope of the Use Case context)	Improve oculavis SHARE network architecture (e.g., shut down server under low load, distribute video calls over infrastructure to reduce minimal server resources needed) and make use of 6Green SBA to decrease operation expenditure to run media server cluster.
Measurement methodology	Comparing operational expenditure of state-of-the-art oculavis SHARE media server cluster compared to improved oculavis SHARE network architecture running in 6Green SBA simulating real usage.
Target	Decreased by 50%
Expected impact on energy efficiency	Reduced energy consumption due to less used server components and lower usage of CPU and RAM.

Table 12: Adaptive bandwidth depending on UE.

5/6G signal quality for UC2 Adaptive bandwidth	KPI_ UC2_3
Description	Based on the given UE 5/6G signal quality, the available bandwidth can fluctuate. In addition to that, an end-user needs a high bandwidth (e.g., for a 4K video stream with 60 fps) on-demand to conduct a remote inspection. The client applications need to adapt its encoding parameter dynamically to work under these changing conditions.
Objective (in scope of the Use Case context)	Client application of oculavis SHARE vApp dynamically adapt to the changing and demanding bandwidth changes by adjusting the video encoding parameters accordingly.
Measurement methodology	Simulating changing bandwidth conditions by throttling network connections and observing/capturing video encoding parameters.
Target	Boolean – Verify the accomplishment by accessing and measuring bandwidth and encoding parameters under real changing or simulated network conditions.
Expected impact on energy efficiency	N.A.

Network latency

Table 13 discusses the latency KPI for UC2. For general considerations on latency and possible measurement setups refer to section 4.2.3.

Table 13: Network Latency KPI for UC2.

Network Latency	KPI_ UC2_4
Description	Network Latency range (Remote Visual Assistance Apps used), Adaptive bandwidth depending on UE 5/6G signal quality
Objective (in scope of the Use Case context)	Remote Visual Assistance application demands specific network latency to provide the supported service with "proper" QoE
Measurement methodology	Latency (in ms) is measured from the client to the central deployment and back ("considerations on the processing time"). Probe based metric collection – client-server-approach. See par. 4.2.3 for further insights.
Target	10-100 ms
Expected impact on energy efficiency	Same considerations as in KPI_UC1_5

4.4 UC3: Zero-Carbon Clientless Virtual Enterprise Desktop as-a-Service

4.4.1 Use Case 3 Description

User story: This is a scenario of remote Desktop as a Service (DaaS) solutions in a real cloud environment. In this use case, end users are connected to a 5G mobile network (which will be totally transparent to the end user) through laptops or mobile phones and will have access to the public Internet. Remote desktop server or servers will be deployed in the central cloud either on a DC provided by some partner of the project, or in a public cloud (such as Amazon EC2). The distance of the remote desktop server and the proximity of the client to the server is one factor that needs to be taken into consideration. The clients will access the remote Desktop through either client (specific binary/apk that will be deployed on the phone/laptops) or through a common browser (able to run HTML5). The clients using **RDP/PCoIP/VNC/TeamViewer/ICA protocol** will access the server either through HTTP or HTTPS. The scenario can be tested with moving end users.

To measure the performance of remote desktop services, there will be three user profiles: **office, web browsing, and video user profiles**. These three profiles present different degrees of interactivity and a varying frequency change to the desktop usage.

The user actions (keyboard presses and mouse movement events) will be recorded using specific capturing tools like the Macro Recorder tool or Touch recorder (for Android phones). The recorded actions will be replayed for each remote desktop environment and each experiment. In this manner, we guarantee the same user actions for all experiments, with the same timing.

The selected metrics for the evaluation of the use case can be:

- **Average power consumption --** obtained by averaging the instantaneous power consumption reported by the meter over the whole duration of an experiment.
    - Other measurable units: Instant Power, Number of Operational units/racks, Server utilization, Number of machines needed, Compute cost per hour, Special HW i.e., FPGA, storage area network (SAN).
- **Network Usage** -- Bandwidth utilization, expressed in terms of the total number of bytes exchanged during an experiment divided by the corresponding duration.
- **Delay** or latency as described in the previous subsection 4.2.3.
- **Packet loss rate**.
- **QoE** -- The service quality experienced by the user of an RD system depends on the video quality and interactivity.
    - QoE is measured based on the difference between the video stream directly at the video output in the source desktop server and at the destination RD client. The **PSNR** will be used as an objective metric of the difference between both video sources.

These metrics are based on the network traffic and video streams at both the server and client.

Objectives: DaaS services can benefit from the 5/6G technological breakthrough: 5/6G low latency allows to shift all the remaining "computational intelligence" from clients (becoming real "zero-clients"/BYOD) to servers, leaving only human I/O tasks on clients. This can further strengthen the potential OpEx and GHG emission gains previously mentioned.

Moreover, passing from the public cloud to 5/6G edge-cloud continuum, DaaS can benefit from 5/6G native security, slice integration with the private enterprise network infrastructure, etc. Therefore, it could become more than attractive for companies and a flexible means to efficiently support employees' smart working (and, therefore, to further reduce GHG emissions).

Architecture: The DaaS architecture is designed with multiple integrated components, which together form the foundation of the application design. The core service is the clientless remote desktop gateway which incorporates support for VNC, RDP, and SSH protocols. This application can be deployed with the infrastructure-as-a-Service (IaaS) such as OpenStack or OpenNebula. In addition, the deployment of all the components will be carried out at the CNIT Testing Laboratory, and the end users (through laptops and UEs) will connect to this deployment.

In Figure 16 it is shown the architecture of the application and the connection between different components.

The architecture includes:

- **Secure Gateway - Front-end**: The front-end component forms the outermost layer of the application, where the remote desktop is rendered to the web browser. This is also the component that implements keyboard, mouse and other events that are sent to the virtual desktop through web socket server and Remote Desktop Protocol Wrapper.
- **Database**: The inclusion of the database is for authentication purposes and to save the user information such as the connection details for a particular user.
- **Web Server:** To communicate with Remote Desktop Protocol Wrapper, there should be an intermediary server that provides a connection between the application and RDP Wrapper. The application is using RDP Wrapper as the proxy for connecting to remote desktops. The server must implement the **guacamole protocol**. The reason behind this is that *guacd* (guacamole proxy daemon)

neither implements nor understands any of the protocol but implements guacamole protocol to understand the connection required from remote desktops.

- **Remote Desktop Protocol Wrapper:** As shown in Figure 16 this component, which is the core component of the application, forms the server-side proxy of the application. The main functionality is to connect to the remote desktops using RDP/VNC/SSH. It keeps the port open for TCP connections incoming from the web applications and detects the type of connection needed for remote connection and connects the user to the appropriate desktop. This acts as a translation layer between the virtual desktops and the web applications.

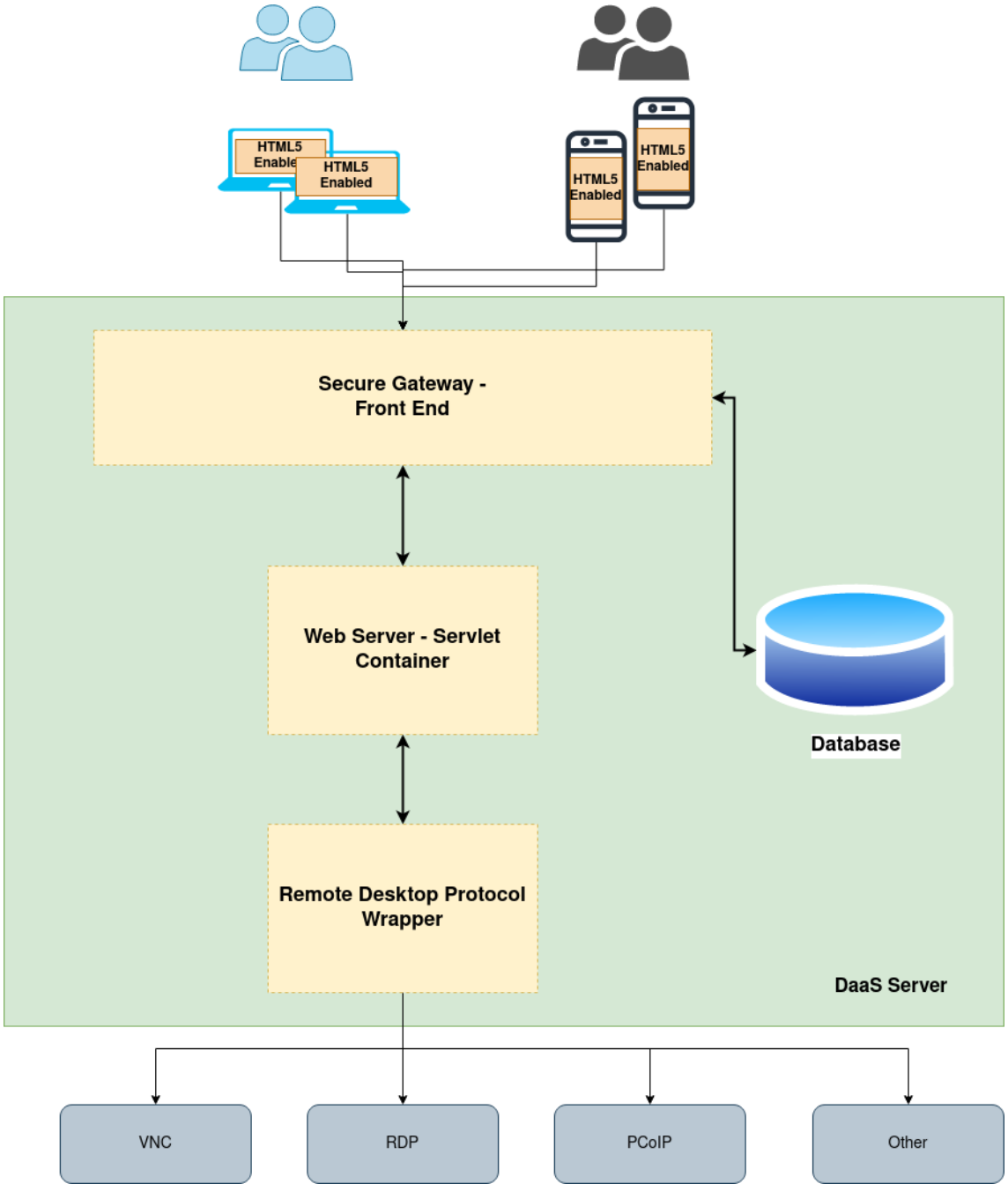

Figure 16: Architecture of the DaaS application.

Business case: With the DaaS paradigm, it is no longer needed to worry about PC hardware costs, and it is possible to instead shift expenses toward an easy to manage subscription service. DaaS also has the added benefit of helping to control IT costs because it requires less in-house technical support to deploy.

DaaS allows companies to provide employees with all necessary rights and permissions in a matter of hours. DaaS is a great option for companies with strapped IT departments and engineering staff that need a quick environment for test and dev.

The (BYOD) bring your own device model creates a unified platform across multiple devices. End devices have very few requirements to become a workspace for every employee. In addition, it enables telecommuters and traveling employees to access information from anywhere, while standardizing performance, security, and support.

Finally, given that in the DaaS model the data is held in the cloud instead of on end-user devices, it is less prone to human errors and security breaches. Providing the capability to deploy a consistent set of user protocols across all devices also fortifies the security from the scope of the Service provider.

Relevant stakeholders: The use case targets infrastructure providers, cloud providers and telecom operators with the goal of enabling for the DaaS paradigm. Specialized technologies such as P4 and GPUs will also be used to boost performance. The 6Green approach will be leveraged to gain insights and assess if the resource intensive DaaS paradigm, from the perspective of the datacenter provider, aligns with the energy criteria proposed by the project.

4.4.2 Use Case 3 Technology/Functional Enablers

Thin or Zero Clients

A DaaS solution offers the ability to deliver virtual desktops and applications to more efficient thin clients or devices that demand less computing power and therefore consume less electricity and produce less heat than their heavier traditional counterparts.

Thin or zero clients are more stripped-down devices, such as some versions of Google Chromebook, which rely on web-based software programs (HTML5 support) or hosted virtual desktops instead of locally installed software. Not only they are generally less expensive than traditional PCs, but thin clients also tend to use less electricity and produce less heat, reducing the need for air conditioning or cooling systems. They also tend to be lighter and smaller than traditional desktop PCs, which means they not only require less raw material to be constructed, but less energy and manufacturing supplies. Their production may also result in less water pollution than larger desktop PCs. The clients will access the remote Desktop through either client (specific binary/apk that will be deployed on the phone/laptops) or through a common browser (able to run HTML5). The clients will access the server either through HTTP or HTTPS, using **RDP/PCoIP/VNC/TeamViewer/ICA protocol**.

Powerful Data centers

Cloud computing minimizes energy needs and therefore consumption by aggregating discrete data centers into large-scale facilities that can more efficiently optimize energy as well as reduce wasted energy (e.g., building facilities in colder climates for natural cooling purposes). By bringing technologies that can virtualize and host multiple virtual Desktop infrastructures allows to significantly lower their energy costs, reduce waste caused by purchasing resource-heavy hardware, and generally ensure more efficient deployment of their budget and resources.

Data centers need to be able to optimize the deployments using enhanced technologies like NIC offloading and optimized CPU and GPU utilization for efficient and fast processing. Optimized DCs are mandatory to be equipped with specific probes for calculating the power consumption and able to distribute the workload across the computing units in order to lower and harmonize the power consumption.

5G Radio

Network latency plays a significant role in the overall user experience in Virtual Desktop Infrastructure (VDI) environments (see Table 16 for insights). Users transmit keystrokes and mouse movements to a remote machine through the network, and the performance of the virtual desktop relies heavily on the quality of network connectivity between the endpoint and VDI. Although VDI has improved in terms of speed, it didn't gain widespread popularity initially due to perceived slowness caused by underlying network issues. However, with the advent of 5G connectivity and support for high-precision time synchronization, it becomes possible to overcome network latency problems and provide a seamless user experience for those operating virtual machines.

4.4.3 Use Case 3 Performance Metrics

Table 14: Carbon reduction, Zero-Carbon Service Agreement KPI.

Carbon reduction, Zero-Carbon Service Agreement	KPI_ UC3_1
Description	Virtual Desktop environments as-a-Service (Desktop as a Service), shifting from thin-client paradigms to zero-client ones. This zero-client environment, DaaS instances are made accessible through dedicated apps that can be installed in any operating systems.
Objective (in scope of the Use Case context)	Less processing power on the client side leads to lower carbon emissions from the laptop/desktop usage.
Measurement methodology	Power, Power/user, Resources (CPU, RAM, NICs), Applications Load can be measured to assess the carbon reduction achieved. A typical desktop and screen used for eight hours results in greenhouse gas (GHG) emissions equivalent to around 70g CO2e arising from the electricity consumed and the Workload.
Target	Decreased by 85%
Expected impact on energy efficiency	Reduced energy consumption due to lower usage of CPU, RAM and Disk

Table 15: Maintenance costs KPI.

Maintenance costs	KPI_ UC3_2
Description	DaaS and zero-client paradigm can reduce the Opex in total (not the infrastructure providers costs)
Objective (in scope of the Use Case context)	Virtual Desktop Infrastructure can reduce Capital Expenses / Operational Expanses more than 85% compared to standard hardware-based solutions
Measurement methodology	The reduction of maintenance cost, which is coupled with the benefits of virtualization, is measured considering, e.g., power, number of operational units/racks, server utilization, number of machines needed, compute cost per hour, special HW i.e., FPGA, storage area network (SAN), keyboard/video/mouse. The measurements can focus on the number of dedicated laptops/desktops (with all the peripherals) that can be replaced with their virtual counterparts in this UC.
Target	Decrease the maintenance cost by 40%
Expected impact on energy efficiency	Reduction of the power used to compare a physical Office (with dedicated devices) with the virtual. On-site/remote maintenance, Cooling System impact, CPU, RAM, Disk

Table 16: Performance-adaptive Network Bandwidth/Latency KPI for UC3.

Performance - Adaptive Network Bandwidth/Latency	KPI_ UC3_3
Description	Network Latency range (depending on vDesktop apps used), adaptive bandwidth depending on UE 5/6G signal quality
Objective (in scope of the Use Case context)	Virtual Desktop application demands specific network latency to provide the supported service with "proper" QoE
Measurement methodology	Latency from the client to the central deployment and back ("considerations on the processing time"). Probe based metric collection – client-server-approach. See par. 4.2.3 for further insights.
Target	15-100 ms
Expected impact on energy efficiency	Same considerations as in KPI_UC1_5

Table 17: Mobility support KPI.

Mobility support	KPI_ UC3_4
Description	vDesktop “migration” according to UE handover
Objective (in scope of the Use Case context)	Virtual Desktop applications can support migration of the user state during handovers of his/her end device
Measurement methodology	Probes that measure the handover period on the telco provider and the QoE/QoS levels
Target	Boolean -- Verify the accomplishment by accessing and measuring a moving end-device to the central VDI deployment in CNIT
Expected impact on energy efficiency	N.A.

Table 18: Zero-client KPI.

Zero-client	KPI_ UC3_5
Description	Create a zero-client environment where DaaS instances are made accessible through dedicated apps that can be installed in any operating systems, or even through native HTML5 interfaces
Objective (in scope of the Use Case context)	Virtual Desktop application will only require a smartphone with 5G connectivity and a Web Browser
Measurement methodology	5G Connectivity KPIs, Web Browser Support. Verification of compatibility with 5G, Web browser ability to run RDP and HTML5
Target	Boolean -- Laptop/mobiles will access the DaaS service through HTML 5 supporting browsers
Expected impact on energy efficiency	Considerations similar to KPI_UC3_2

5 Conclusions

In accordance with the objectives of WP2, this document serves as an initial introduction to the project, providing an explanation of the main research areas and innovations while clarifying the ultimate goals.

It presents a generic Service-Based Architecture (SBA) framework to identify the key architectural elements and technological enablers that form the foundation of the 6Green ecosystem. Furthermore, it includes a comprehensive list of general 6G network requirements, which are analysed to outline the context in which the specific 6Green platform will be developed.

The second part of the document focuses on the analysis of relevant use cases used as benchmarks to assess the performance of the proposed 6Green solution. For each scenario under study, the document provides a general description that emphasizes the architectural and technological elements encompassed. Additionally, it includes a list of metrics and Key Performance Indicators (KPIs) that are relevant to the specific use case, as well as some insights into the business case and stakeholders involved.

When it comes to KPIs, the document goes beyond identifying traditional performance metrics and target parameters. It aims to define innovative KPIs that surpass the current State-of-the-Art (SoTA) and are tailored specifically for the 6Green ecosystem. These KPIs will be used in other technical Work Packages (WPs) to assess the performance of the 6Green platform and evaluate its impact on energy consumption across relevant scenarios.

In conclusion, this first deliverable plays a crucial role in paving the way for all future project tasks. Its primary objective is to establish clear project guidelines, pinpointing the key technological enablers and significant innovations that form the foundation of the 6Green platform. Additionally, it involves carefully selecting a range of use cases to effectively demonstrate the platform's value within the realm of sustainable 6G-based mobile technologies.

References

[128.533] ETSI TS 128 533, "5G; Management and Orchestration; Architecture framework", version 17.2.0, Rel. 17, March 2022.

[23.288] 3GPP. TS 23.288 Architecture Enhancements for 5G System to Support Network Data Analytics Services, version 16.5.0; 3GPP: Sophia Antipolis, France, 2020.

[23.501] 3GPP, TS 23.501, "System architecture for the 5G System (5GS) (Release 17)", 2023-04.

[23.502] 3GPP, TS 23.502, "Procedures for the 5G System (5GS), (Release 17)", 2023-04.

[23.503] 3GPP, TS 23.503, "Policy and charging control framework for the 5G System (5GS); Stage 2 (Release 17)", 2023-04.

[23.971] 3GPP TR 23.791 "Study of enablers for network automation for 5G (release 16)," June 2019.

[28.533] 3GPP TS 28.533.: 3rd Generation Partnership Project; Technical Specification Group Services and System Aspects; Management and Orchestration; Architecture Framework; Stage 3 (Release 16), V16.3.0., March 2020.

[29.501] 3GPP TS 29.501 "5G System; Principles and Guidelines for Services Definition; Stage 3," September 2020.

[5GI21] https://www.5g-induce.eu/.

[5GPPP22] "The 6G Architecture Landscape", 5G PPP Architecture Working Group, Version 1.0, December 2022.

[ARZ+22] F. Adinegoro, C. Rahmania, I. N. Zaini, R. M. Negara and S. N. Hertiana, "Latency and RAM Usage Comparison of Advanced and Lightweight Service Mesh," 2022 5th International Seminar on Research of Information Technology and Intelligent Systems (ISRITI), Yogyakarta, Indonesia, 2022, pp. 369-372, doi: 10.1109/ISRITI56927.2022.10052905.

[ASS20] Ali, S., Saad, W., & Steinbach, D. (Eds.). (2020). "White Paper on Machine Learning in 6G Wireless Communication Networks" [White paper]. (6G Research Visions, No. 7). University of Oulu. http://urn.fi/ urn:isbn:9789526226736.

[Bha21] Vivek M. Bhasi, Jashwant Raj Gunasekaran, Prashanth Thinakaran, Cyan Subhra Mishra, Mahmut Taylan Kandemir, and Chita Das. "Kraken: Adaptive Container Provisioning for Deploying Dynamic DAGs in Serverless Platforms". In Proceedings of the ACM Symposium on Cloud Computing (SoCC '21).

[Che23] Chen Chen, Lars Nagel, Lin Cui, and Fung Po Tso. 2023. "S-Cache: Function Caching for Serverless Edge Computing." In Proceedings of the 6th International Workshop on Edge Systems, Analytics and Networking (EdgeSys '23).

[Cic22a] C. Cicconetti, M. Conti and A. Passarella, "In-Network Computing With Function as a Service at the Edge," in Computer, vol. 55, no. 9, pp. 65-73, Sept. 2022.

[Cic22b] C. Cicconetti, M. Conti, A. Passarella, "FaaS execution models for edge applications", Pervasive and Mobile Computing, Volume 86, 2022, 101689, ISSN 1574-1192.

[CKB22] H. Chergui, A. Ksentini, L. Blanco and C. Verikoukis, "Toward Zero-Touch Management and Orchestration of Massive Deployment of Network Slices in 6G," in IEEE Wireless Communications, vol. 29, no. 1, pp. 86-93, February 2022, doi: 10.1109/MWC.009.00366.

[CMS22] A. Chouman, D. M. Manias and A. Shami, "An NWDAF Approach to 5G Core Network Signaling Traffic: Analysis and Characterization," GLOBECOM 2022–- 2022 IEEE Global Communications Conference, Rio de Janeiro, Brazil, 2022, pp. 6001-6006, doi: 10.1109/GLOBECOM48099.2022.10000989.

[CMS22b] A. Chouman, D. M. Manias and A. Shami, "Towards Supporting Intelligence in 5G/6G Core Networks: NWDAF Implementation and Initial Analysis," 2022 International Wireless Communications and Mobile Computing (IWCMC), Dubrovnik, Croatia, 2022, pp. 324-329, doi: 10.1109/IWCMC55113.2022.9824403.

[CNF17] Cloud Native Computing Foundation, https://www.cncf.io/.

[Cos22] B. Costa et al., "Orchestration in fog computing: A comprehensive survey," ACM Comput. Surv., vol. 55, no. 2, jan 2022.

[DCD22] S. Dustdar, V. Casamajor Pujol, and P. K. Donta, "On distributed computing continuum systems," IEEE Transactions on Knowledge and Data Engineering, pp. 1–1, 2022.

[DK23] V. -B. Duong and Y. Kim, "A Design of Service Mesh Based 5G Core Network Using Cilium," 2023 International Conference on Information Networking (ICOIN), Bangkok, Thailand, 2023, pp. 25-28, doi: 10.1109/ICOIN56518.2023.10049044.

[EGD19] D. M. Gutierrez-Estevez et al., "Artificial Intelligence for Elastic Management and Orchestration of 5G Networks," in IEEE Wireless Communications, vol. 26, no. 5, pp. 134-141, October 2019, doi: 10.1109/MWC.2019.1800498.

[ENI005] ETSI GS ENI 005 v1.1.1 "Experiential networked intelligence (eni); system architecture," 2019.

[ENI012] "Experiential Networked Intelligence (ENI) Group Report", URL: https://www.etsi.org/deliver/etsi_gr/ENI/001_099/012/01.01.01_60/gr_ENI012v010101p.pdf.

[FOS23] Emily Foster, "Kepler project aims to help curb Kubernetes energy waste", https://www.techtarget.com/searchitoperations/feature/How-to-approach-sustainability-in-IT-operations.

[Fu22] K. Fu et al., "Adaptive resource efficient microservice deployment in cloud-edge continuum," IEEE Transactions on Parallel and Distributed Systems, vol. 33, no. 8, pp. 1825–1840, 2022.

[GRR+19] E. García-Martín, C. F. Rodrigues, G. Riley, H. Grahn, "Estimation of energy consumption in machine learning," Journal of Parallel and Distributed Computing, Volume 134, 2019, Pages 75-88, ISSN 0743-7315, https://doi.org/10.1016/j.jpdc.2019.07.007.

[HCC22] Hsiung, C., Lin, F.J., Chen, JC., Chen, C. (2022), "5G Network Slice Scalability Based on Management Data Analytics Function (MDAF)," In: Hsieh, SY., Hung, LJ., Klasing, R., Lee, CW., Peng, SL. (eds) New Trends in Computer Technologies and Applications. ICS 2022. Communications in Computer and Information Science, vol 1723. Springer, Singapore. https://doi.org/10.1007/978-981-19-9582-8_52.

[Heu21] Martijn de Heus, Kyriakos Psarakis, Marios Fragkoulis, and Asterios Katsifodimos. "Distributed transactions on serverless stateful functions". In Proceedings of the 15th ACM International Conference on Distributed and Event-based Systems (DEBS '21).

[IFA013] Network Functions Virtualisation (NFV) Release 3; Management and Orchestration; Os-Ma-nfvo reference point – Interface and Information Model Specification, https://www.etsi.org/deliver/etsi_gs/NFV-IFA/001_099/013/03.04.01_60/gs_NFV-IFA013v030401p.pdf.

[ITUML5G] https://www.itu.int/pub/T-FG-ML5G-2019.

[Jin21] Z. Jin, Y. Zhu, J. Zhu, D. Yu, C. Li, R. Chen, I.E. Akkus, Y. Xu, Lessons learned from migrating complex stateful applications onto serverless platforms, in: ACM APSys 2021.

[JM20] L. J. Jagadeesan and V. B. Mendiratta, "When Failure is (Not) an Option: Reliability Models for Microservices Architectures," 2020 IEEE International Symposium on Software Reliability Engineering Workshops (ISSREW), Coimbra, Portugal, 2020, pp. 19-24, doi: 10.1109/ISSREW51248.2020.00031.

[JSZ+21] K. Jiang, C. Sun, H. Zhou, X. Li, M. Dong and V. C. M. Leung, "Intelligence-Empowered Mobile Edge Computing: Framework, Issues, Implementation, and Outlook," in IEEE Network, vol. 35, no. 5, pp. 74-82, September/October 2021, doi: 10.1109/MNET.101.2100054.

[Kat22] Christos Katsakioris, Chloe Alverti, Vasileios Karakostas, Konstantinos Nikas, Georgios Goumas, and Nectarios Koziris." FaaS in the age of (sub-)µs I/O: a performance analysis of snapshotting." In Proceedings of the 15th ACM International Conference on Systems and Storage (SYSTOR '22).

[KEP22] Kubernetes Efficient Power Level Exporter (Kepler), https://sustainable-computing.io/.

[Kim21] D. Kimovski et al., "Cloud, fog, or edge: Where to compute?" IEEE Internet Computing, vol. 25, no. 4, pp. 30–36, 2021.

[KNV+18] M. G. Kibria, K. Nguyen, G. P. Villardi, O. Zhao, K. Ishizu and F. Kojima, "Big Data Analytics, Machine Learning, and Artificial Intelligence in Next-Generation Wireless Networks," in IEEE Access, vol. 6, pp. 32328-32338, 2018, doi: 10.1109/ACCESS.2018.2837692.

[Kok22] H. Kokkonen et al., "Autonomy and intelligence in the computing continuum: Challenges, enablers, and future directions for orchestration," 2022. [Online]. Available: https://arxiv.org/abs/2205.01423.

[LBK+23] S. Lagen, B. Bojovic, K. Koutlia, X. Zhang, P. Wang and Q. Qu, "QoS Management for XR Traffic in 5G NR: A Multi-Layer System View & End-to-End Evaluation," in IEEE Communications Magazine 2023.

[LS22] Kaiyi Liu, Paul Singh, "Sustainability in Computing," https://github.com/sustainable-computing-io/kepler-doc/blob/450c70ae5de2f20238158c8a156daf28f4e5cd51/demos/Kubernetes-Edge-Day.pdf

[Man18] J. Manner, M. Endreß, T. Heckel and G. Wirtz, "Cold Start Influencing Factors in Function as a Service," 2018 IEEE/ACM International Conference on Utility and Cloud Computing Companion (UCC Companion), 2018, pp. 181-188.

[MAT17] https://5g-ppp.eu/matilda/.

[MCD22] F. Malandrino, C. F. Chiasserini and G. Di Giacomo, "Energy-efficient Training of Distributed DNNs in the Mobile-edge-cloud Continuum," 2022 17th Wireless On-Demand Network Systems and Services Conference (WONS), Oppdal, Norway, 2022, pp. 1-4, doi: 10.23919/WONS54113.2022.9764487.

[MCS22] D. M. Manias, A. Chouman and A. Shami, "An NWDAF Approach to 5G Core Network Signaling Traffic: Analysis and Characterization," GLOBECOM 2022 - 2022 IEEE Global Communications Conference, Rio de Janeiro, Brazil, 2022, pp. 6001-6006, doi: 10.1109/GLOBECOM48099.2022.10000989.

[NGMN23] 6G REQUIREMENTS AND DESIGN CONSIDERATIONS by NGMN Alliance e.V, Feb 2023

[NMM+16] Irakli Nadareishvili, Ronnie Mitra, Matt McLarty, and Mike Amundsen. 2016, “Microservice Architecture: Aligning Principles, Practices, and Culture (1st. ed.),” O'Reilly Media, Inc.

[NZS+22] Y. Niu, S. Zhao, X. She and P. Chen,“"A Survey of 3GPP Release 18 on Network Data Analytics Function Management”" 2022 IEEE/CIC International Conference on Communications in China (ICCC Workshops), Sanshui, Foshan, China, 2022, pp. 146-151, doi: 10.1109/ICCCWorkshops55477.2022.9896472.

[ONA17] Open Network Automation Platform, https://www.onap.org/.

[OSM16] Open Source MANO, https://osm.etsi.org/wikipub/index.php/Main_Page.

[Pat21] P. Patros, J. Spillner, A. V. Papadopoulos, B. Varghese, O. Rana and S. Dustdar, “Toward Sustainable Serverless Computing,” in IEEE Internet Computing, vol. 25, no. 6, pp. 42-50, 1 Nov.-Dec. 2021.

[PBC+20] Peltonen, E., Bennis, M., Capobianco, M., Debbah, M., Ding, A., Gil-Castiñeira, F., Jurmu, M., Karvonen, T., Kelanti, M., Kliks, A., Leppänen, T., Lovén, L., Mikkonen, T., Rao, A., Samarakoon, S., Seppänen, K., Sroka, P., Tarkoma, S., & Yang, T. (2020). “6G White Paper on Edge Intelligence” [White paper]. (6G Research Visions, No. 8). University of Oulu. http://urn.fi/urn:isbn:9789526226774.

[Rau21] T. Rausch, A. Rashed and S. Dustdar, “Optimized container scheduling for data-intensive serverless edge computing”, Future Generation Comput. Syst., vol. 114, pp. 259-271, Aug. 2021.

[Ros22] D. Rosendo et al., “Distributed intelligence on the edge-to-cloud continuum: A systematic literature review,” Journal of Parallel and Distributed Computing, vol. 166, pp. 71–94, 2022.

[SC16] Singh, S., Chana, I., “A Survey on Resource Scheduling in Cloud Computing: Issues and Challenges,” J Grid Computing 14, 217–264 (2016). https://doi.org/10.1007/s10723-015-9359-2.

[SCA20] Scaphandre documentations https://github.com/hubblo-org/scaphandre.

[SCA20b] Scaphandre documentation, https://hubblo-org.github.io/scaphandre-documentation/explanations/how-scaph-computes-per-process-power-consumption.html

[Sch21] J. Schleier-Smith et al., “What serverless computing is and should become: The next phase of cloud computing”, Commun. ACM, vol. 64, no. 5, pp. 76-84, 2021.

[SDA+20] Saarnisaari, H., Dixit, S., Alouini, M.-S., Chaoub, A., Giordani, M., Kliks, A., Matinmikko-Blue, M., & Zhang, N. (Eds.). (2020). “6G White Paper on Connectivity for Remote Areas” [White paper]. (6G Research Visions, No. 5). University of Oulu. http://urn.fi/urn:isbn:9789526226750.

[Smi21] F. Smirnov et al., “Apollo: Towards an efficient distributed orchestration of serverless function compositions in the cloud-edge continuum,” in Proceedings of the 14th IEEE/ACM International Conference on Utility and Cloud Computing, ser. UCC ’21. New York, USA: Association for Computing Machinery, 2021.

[Sre20] Vikram Sreekanti, Chenggang Wu, Xiayue Charles Lin, Johann Schleier-Smith, Joseph E. Gonzalez, Joseph M. Hellerstein, and Alexey Tumanov. “Cloudburst: stateful functions-as-a-service”. Proc. VLDB Endow. 13, 12 (August 2020), 2438–2452.

[TDM20] O. Tomarchio, D. Calcaterra, and G. D. Modica, “Cloud resource orchestration in the multi-cloud landscape: a systematic review of existing frameworks,” Journal of Cloud Computing, vol. 9, no. 1, p. 49, Sep. 2020.

[TSM+21] H. Tataria, M. Shafi, A. F. Molisch, M. Dohler, H. Sjöland and F. Tufvesson, "6G Wireless Systems: Vision, Requirements, Challenges, Insights, and Opportunities," in Proceedings of the IEEE, vol. 109, no. 7, pp. 1166-1199, July 2021, doi: 10.1109/JPROC.2021.3061701.

[Tze21] Tzenetopoulos, A. et al. (2021). FaaS and Curious: Performance Implications of Serverless Functions on Edge Computing Platforms. In: Jagode, H., Anzt, H., Ltaief, H., Luszczek, P. (eds) High Performance Computing. ISC High Performance 2021. Lecture Notes in Computer Science, vol 12761. Springer, Cham.

[VIZ23] Mike Vizard, “Red Hat Donates Kepler Tool for Tracking Power Usage CNCF”, https://cloudnativenow.com/features/red-hat-donates-kepler-tool-for-tracking-power-usage-to-cncf/.

[WBB+20] M. Wurster, U. Breitenbücher, A. Brogi, F. Leymann, and J. Soldani, “Cloud-native deploy-ability: An analysis of required features of deployment technologies to deploy arbitrary cloudnative applications,” in Proc. 10th Int. Conf. Cloud Comput. Services Sci., 2020, pp. 171-180.

[WGN+19] S. Wang, Y. Guo, N. Zhang, P. Yang, A. Zhou and X. Shen, “Delay-Aware Microservice Coordination in Mobile Edge Computing: A Reinforcement Learning Approach,” in IEEE Transactions on Mobile Computing, vol. 20, no. 3, pp. 939-951, 1 March 2021, doi: 10.1109/TMC.2019.2957804.

[WML+22] Y. -T. Wang, S. -P. Ma, Y. -J. Lai and Y. -C. Liang, “Analyzing and Monitoring Kubernetes Microservices based on Distributed Tracing and Service Mesh,” 2022 29th Asia-Pacific Software Engineering Conference (APSEC), Japan, 2022, pp. 477-481, doi: 10.1109/APSEC57359.2022.00066.

[WRE20] Y. Wang et al., “From Design to Practice: ETSI ENI Reference Architecture and Instantiation for Network Management and Orchestration Using Artificial Intelligence,” in IEEE Communications Standards Magazine, vol. 4, no. 3, pp. 38-45, September 2020, doi: 10.1109/MCOMSTD.001.1900039.

[YAX+20] H. Yang, A. Alphones, Z. Xiong, D. Niyato, J. Zhao and K. Wu, “Artificial-Intelligence-Enabled Intelligent 6G Networks,” in IEEE Network, vol. 34, no. 6, pp. 272-280, November/December 2020, doi: 10.1109/MNET.011.2000195.

[ZSM002] ETSI GS ZSM 002: "Zero-touch Network and Service Management (ZSM); Reference Architecture V.1.1 (2019-08)".

[ZYH+20] H. Zhang, N. Yang, W. Huangfu, K. Long and V. C. M. Leung, "Power Control Based on Deep Reinforcement Learning for Spectrum Sharing," in IEEE Transactions on Wireless Communications, vol. 19, no. 6, pp. 4209-4219, June 2020, doi: 10.1109/TWC.2020.2981320.