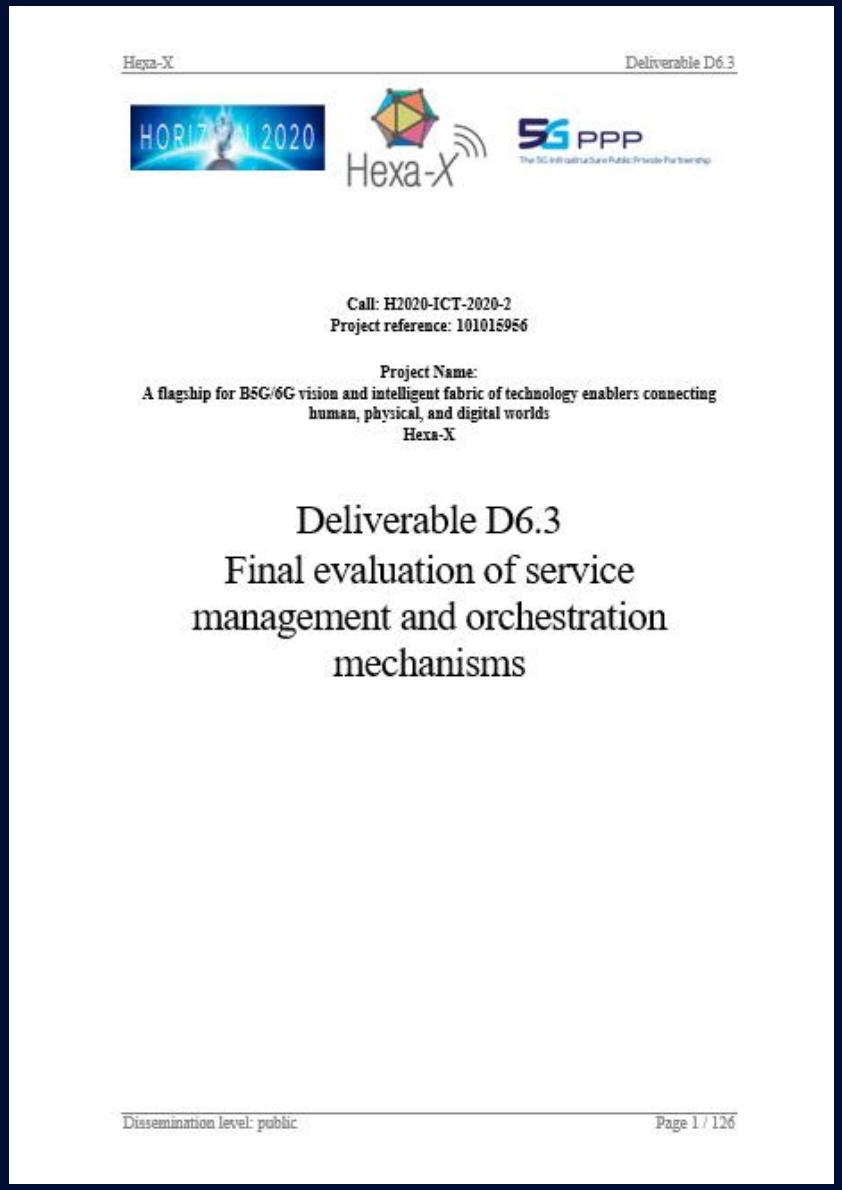
Hexa-X: WP6 – Deliverable D6.3

Final evaluation of service management and orchestration mechanisms

6.04.2023

hexa-x.eu

Mission and Scope



- The deliverable serves as a means of verification of the project Objective 3 (Connecting intelligence towards 6G) and evaluates the benefits of the proposed framework and AI-driven mechanisms and their contribution to project quantifiable targets
- This deliverable introduces the implementation of novelties described in deliverable D6.2, such as
  - Unified orchestration across the “extreme-edge, edge, core” continuum
  - Unified management and orchestration across multiple domains owned and administered by different stakeholders
  - Increasing levels of automation
  - adoption of data-driven and AI/ML techniques in the M&O system and
  - adoption of the cloud-native principles in the telco-grade environment
- This implementation takes form of two demos and a set of complementary lab experiments

Outline

Overview

Demo #4:
Handling unexpected situations in industrial contexts

Demo #5:
Data-driven device-edge-cloud continuum management

Complementary lab Experiments:

- Network Energy Efficiency
- Extreme-Edge nodes discovery
- Simu5G in Scenario 5.1

Evaluation

Use Cases alignment

- Demo #4 is in strong alignment with two use cases described in D1.2, namely:
  - Digital Twins for manufacturing, and
  - Flexible manufacturing.

  Both of these use cases are related to each other. The first one describes how using Digital Twins can benefit production lines via improvements in capabilities such as management of infrastructure resources, detection of anomalous behaviour, and mitigation of critical situations. The second one focuses on allowing dynamic configuration of real-time communication services, which is essential for mobile production machinery.

- Demo #5 is aligned with one described in D1.2 use case, which is **6G IoT micro-networks for smart cities**. This use case focuses on the management of traffic flows in a complex local system of objects interacting with each other. Traffic Light control described in Demo #5 can be such a system.

Work Topics and targeted innovations

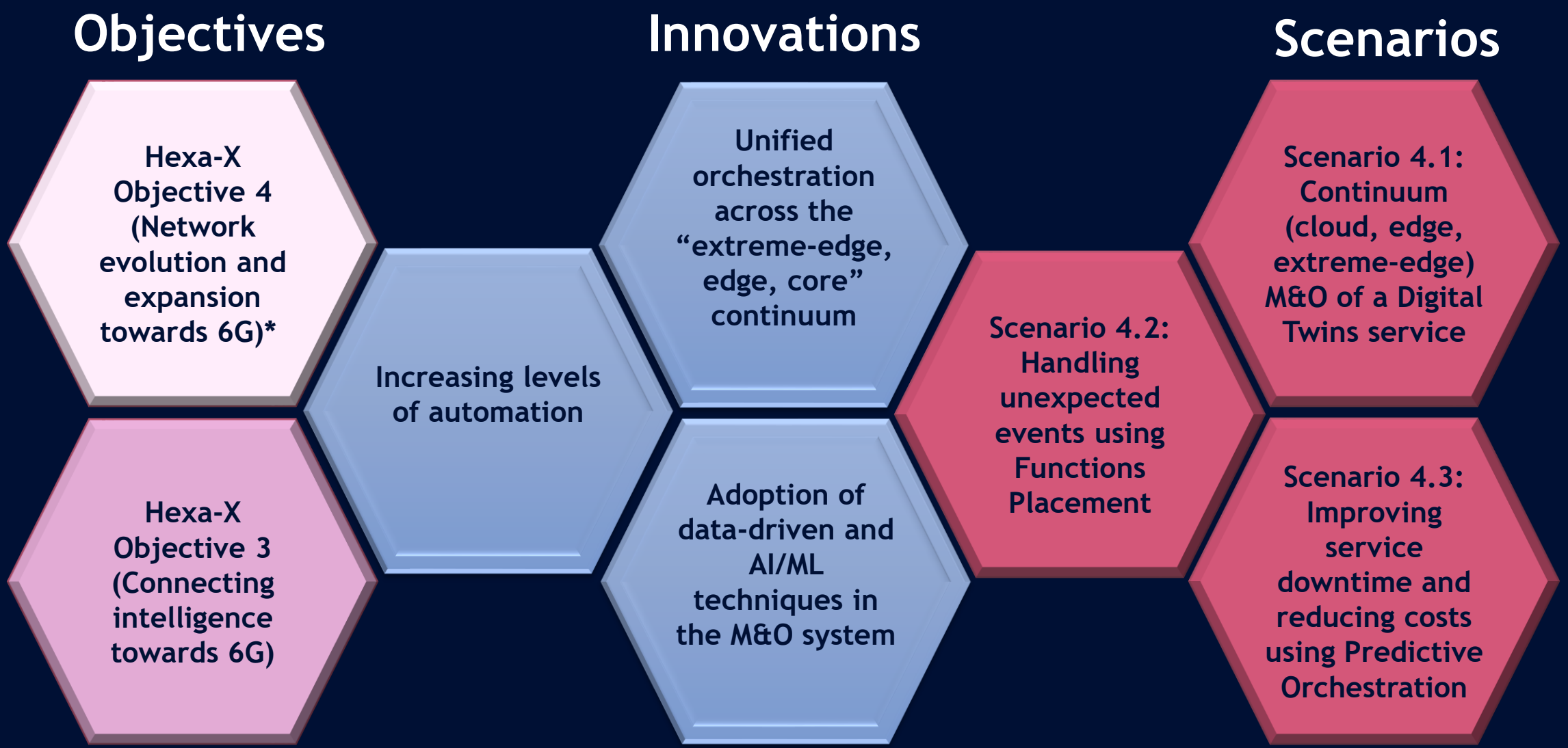
Hexa-X

Demos and lab experiments in this Delivarable cover the following work topics and innovations posed in the previous Deliverable D6.2:

Work Topic	Demo #4	Demo #5	Lab Exp.
1: Unified orchestration across the “extreme-edge, edge, core” continuum	✓	✓	✓
2: Increased level of automation	✓	✓	
3: Adoption of data-driven and AI/ML techniques in the M&O system	✓	✓	
4: Unified management and orchestration across multiple domains, owned and administered by different stakeholders		✓	
5: Adoption of the cloud-native principles in the telco-grade environment		✓	
6: Security		✓	
7: RAN integration			✓
8: Network energy efficiency			✓

Demo #4: Handling unexpected situations in industrial contexts

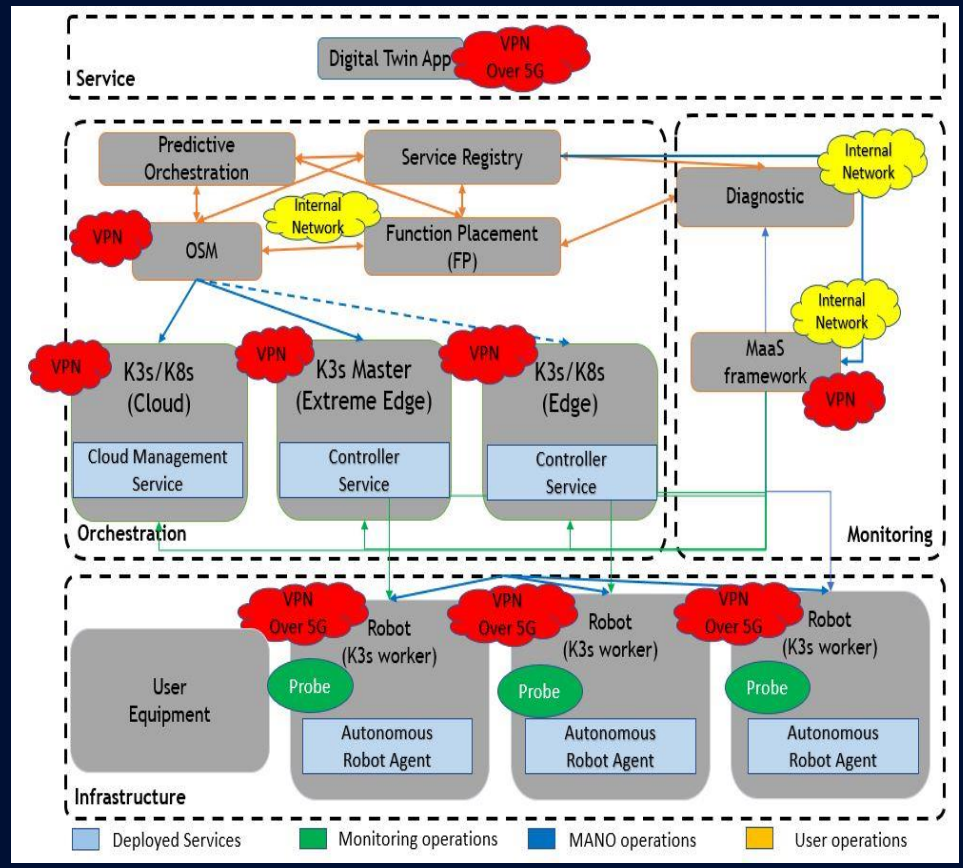
Demo #4 - Overview



**Objective 4 is WP7 objective. Included here because Demo #4 is a joint WP6-WP7 demo*

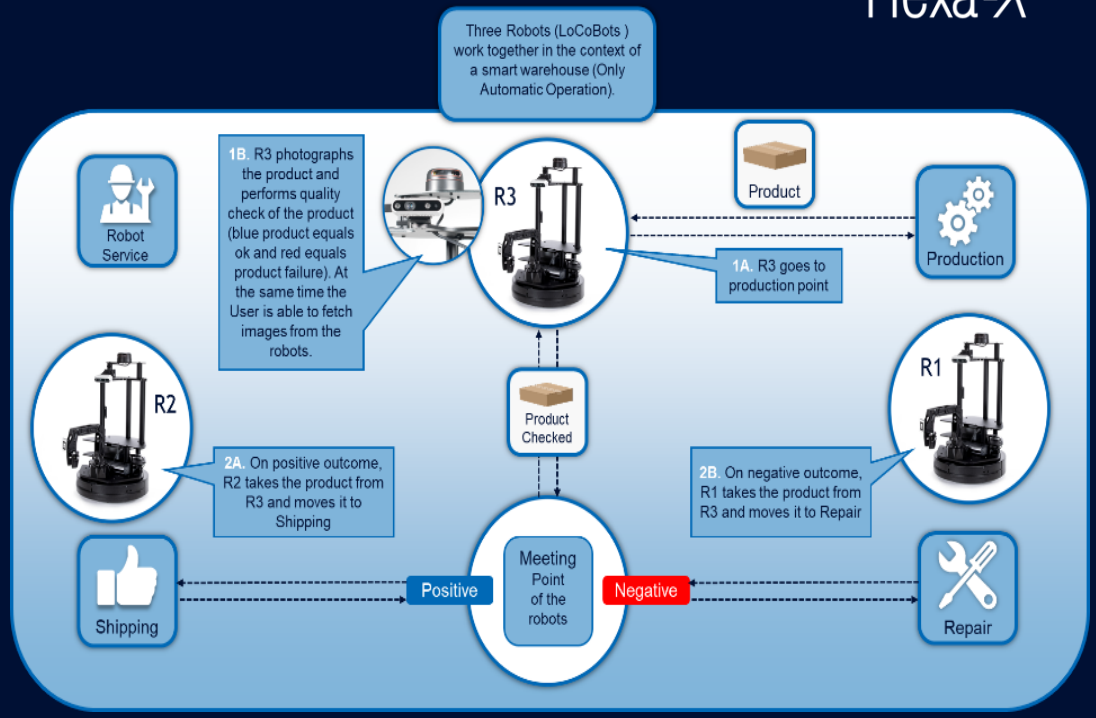
Demo #4 - Scenario 4.1: Continuum (cloud, edge, extreme-edge) M&O of a Digital Twins service

- 3 mobile robots in a simulated industrial environment, working as *Digital Twins*
- Robots can "copy" and execute the actions and movements executed by humans through a remote Human-Machine Interface (HMI)
- This avoids the human presence in the industrial environment itself, which could be inconvenient or even dangerous in some cases
- The HMI has been implemented through an advanced Virtual Reality User Interface (UI), which can be used to visualize the whole industrial environment in real-time and with 3D graphics
- The UI can be used also used to monitor the robots themselves, providing a great level of detail
- The end-user can manage the robots, detect issues, and fix them through teleoperation



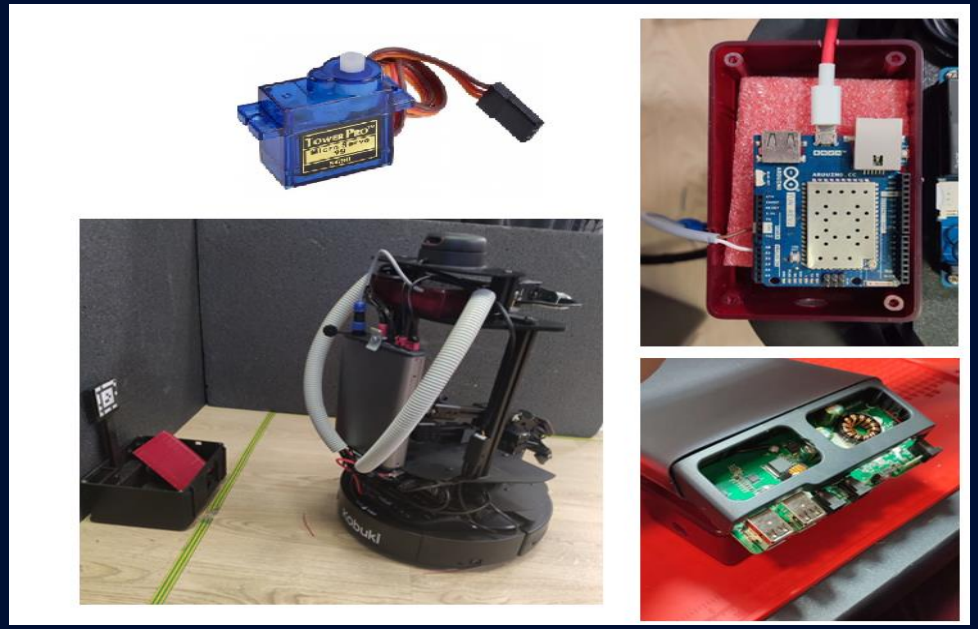
Demo #4 - Scenario 4.2: Handling unexpected events using Functions Placement

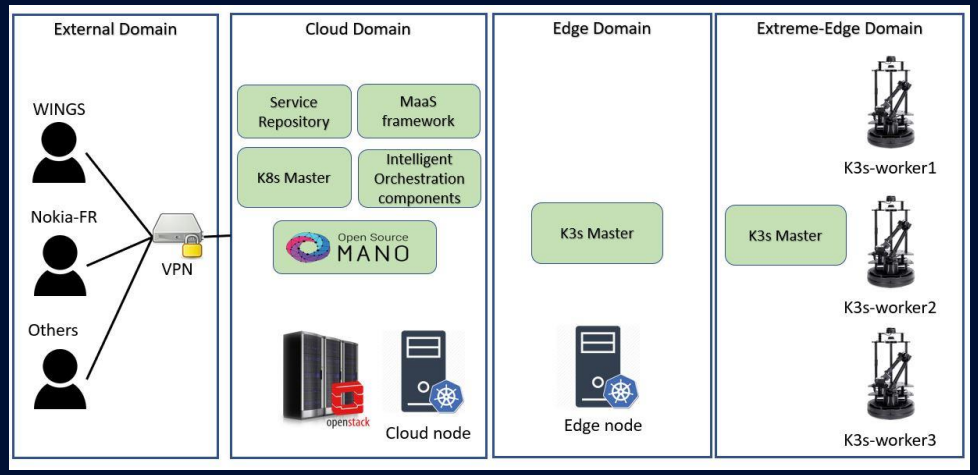
- Similarly to Scenario 4.1, an emulated industrial production line has been built using three mobile robots, with three locations assigned as goals for their respective roles of the robots and placeholder objects that are to be transferred between the different target locations: Production (quality checking), Shipping, and Repairing, based on the role that each robot implements
- This scenario aims to demonstrate how AI/ML enablers, for anomaly detection and performance degradation analysis, along with increased automation and programmability, can be utilized to further increase the efficiency of network and/or service operations, in the simulated industrial context, using closed-loop control mechanisms
- These mechanisms rely on monitoring and performance diagnosis of the various services and components running on the infrastructure and are responsible for reconfiguring and redeploying services and functionalities in order to optimize their performance and achieve the targeted KPIs/KVIs
- The Function Placement component (which is described in detail in Deliverable D7.3, sections 6.1.4 and 6.1.5) is developed with the purpose of optimizing the placement of services and their components across the available infrastructure, either at the cloud, edge or the extreme-edge domains



Demo #4 - Scenario 4.3:
Improving service downtime and reducing costs using Predictive Orchestration

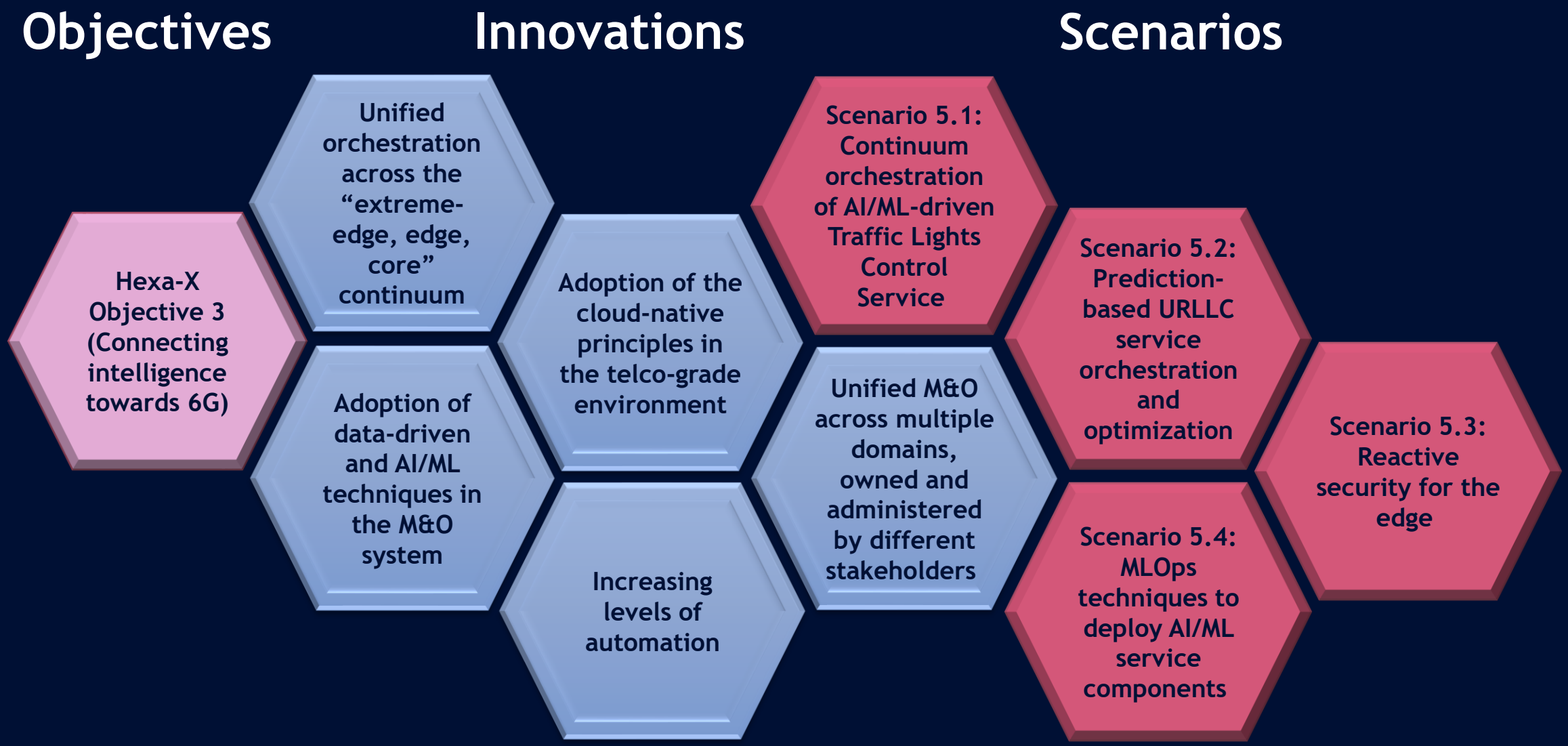
- This scenario demonstrated how AI/ML enablers for prediction of the behaviour of services and/or components can lead to increased efficiency and reduced costs of industrial operations and systems. The focus was on the health and power consumption of the robots, like their batteries and motors (top image)
- The monitoring data from selected services/components have been used to train predictive AI/ML model to identify upcoming critical events, such as malfunctions, overvoltage, extreme stress, low power, etc., and trigger the appropriate orchestration actions to avoid/handle them pre-emptively
- This scenario utilizes all the software components introduced in the two previous scenarios and extends them with the introduction of the predictive orchestration component





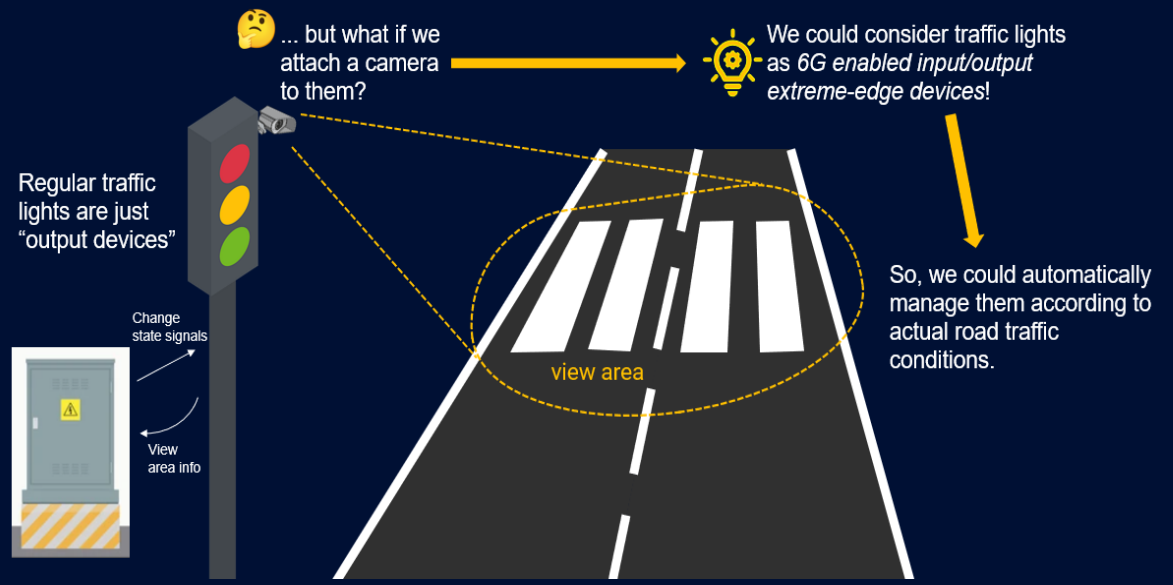
Demo #5: Data-driven device-edge-cloud continuum management

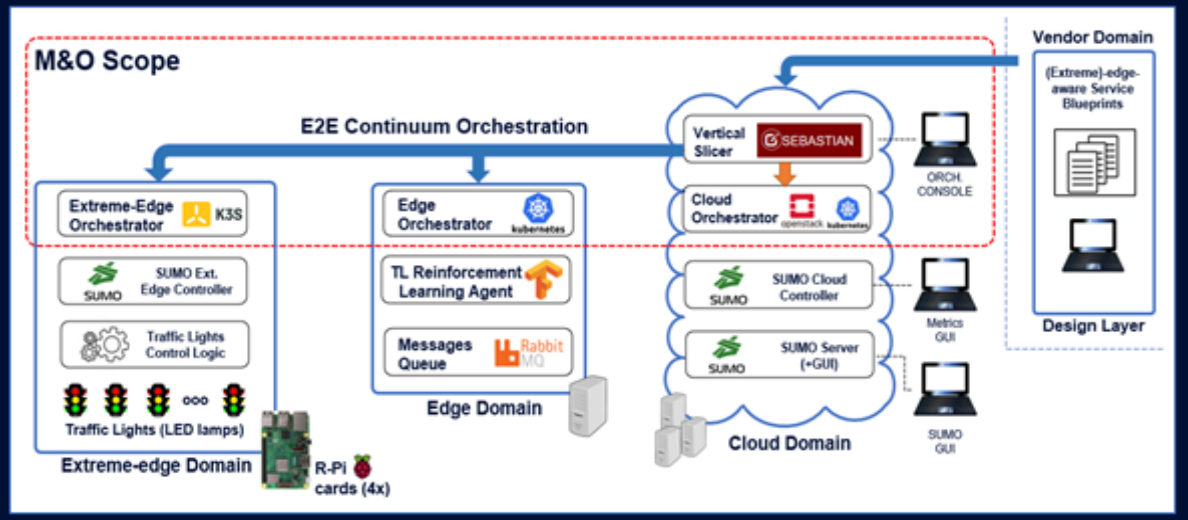
Demo #5 - Overview



Demo #5 - Scenario 5.1:
Continuum orchestration of AI/ML-driven Traffic Lights Control Service

- This scenario aims at demonstrating how the 6G technology can be used to improve the road traffic flow in urban environments by controlling the traffic lights using AI/ML
- The objective is to demonstrate how the deployment of this AI/ML-enabled control could improve road traffic mobility compared to the common approach, i.e., the traffic lights activation based only on periodic time patterns
- The scenario considers more advanced traffic lights that, beyond simply switching on/off their lamps, would also be enabled to perceive their immediate environment (e.g., through cameras installed on the traffic lights themselves or in nearby locations – see Figure)
- This surrounding information will be sent to an AI/ML application at the edge through the 6G network, where it will be processed to trigger more intelligent actions by adapting the traffic lights switching times to the actual traffic conditions
- This would help to reduce traffic jams and minimize waiting times. A real-life implementations of a system like this could help to reduce CO_2 emissions in polluted cities
- This AI/ML service is orchestrated through the network *continuum*, with elements at the extreme-edge (the traffic lights themselves and their controllers), the edge (the AI/ML agents), and the central cloud



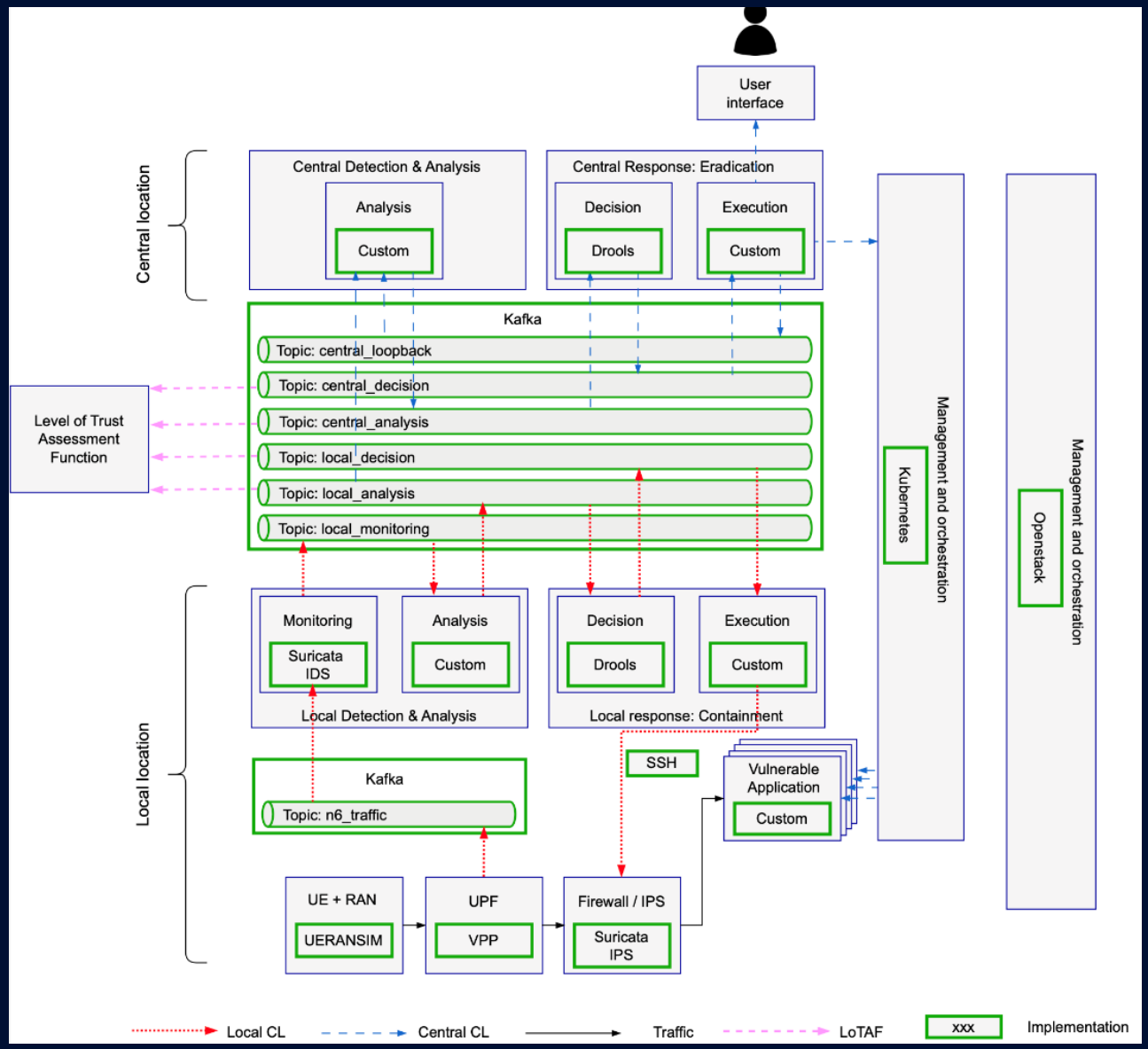


Demo #5 - Scenario 5.2:
Prediction-based URLLC service orchestration and optimization

- This scenario aims at demonstrating the ability of machine learning algorithms to anticipate the resource needs of the network and pre-emptively activate the related services so the application perceives no delay
- In a nutshell, in contrast to reactive methods that may scale up/down the resources as the traffic load increases/decreases, this scenario demonstrates the advantages of a proactive approach that does not involve the typical delays of reactive methods
- This is particularly critical for deployments where resources are set in a deactivated or sleep state to support sustainability but require a non-negligible amount of time to be powered on. Such boot-up delays are highly harmful in the case of real-time services, such as URLLC services

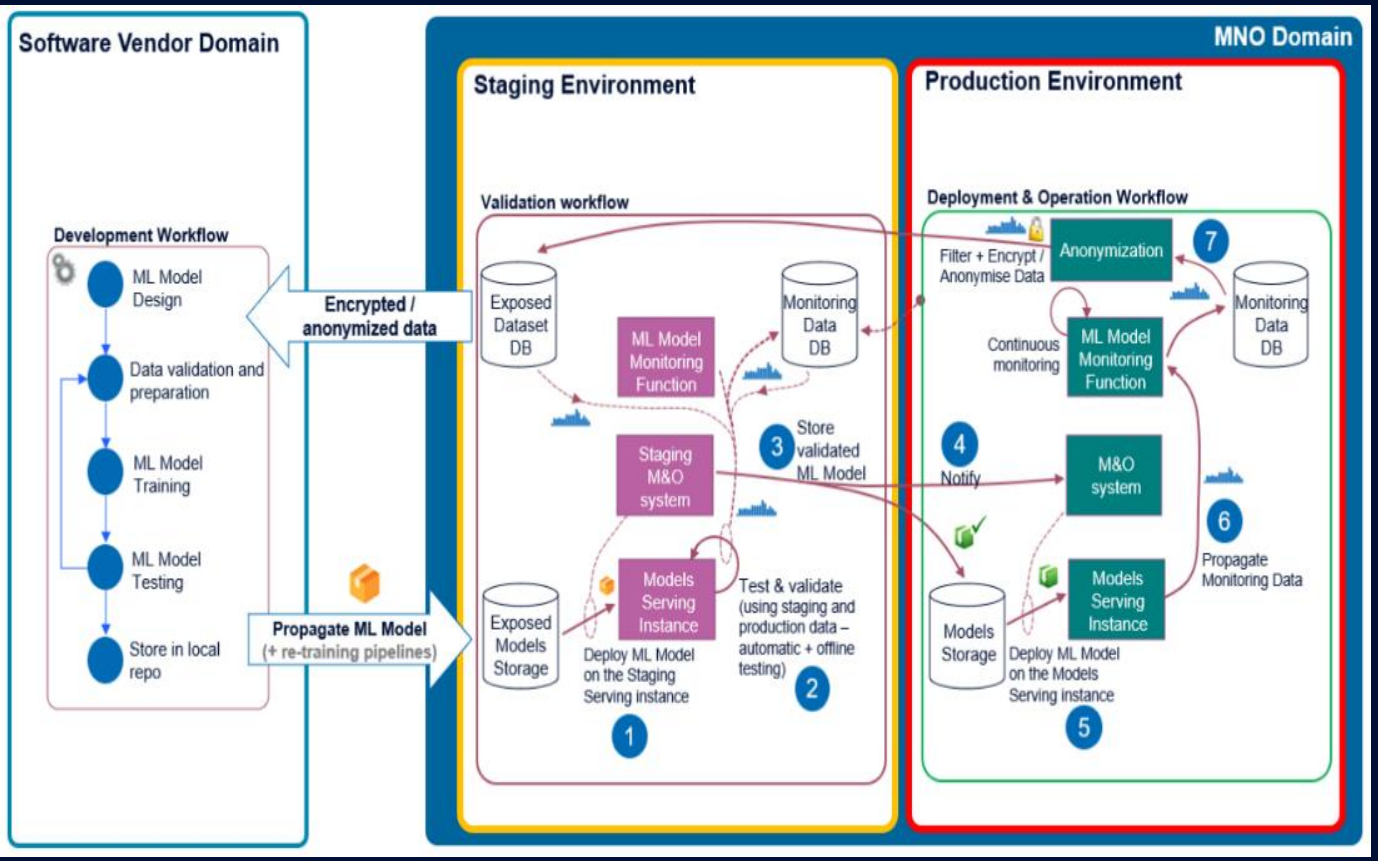
Demo #5 - Scenario 5.3: Reactive security for the edge

- The use cases developed in the previous Scenarios 5.1 and 5.2 regarding road traffic related applications, involve resources and services deployed over the extreme-edge
- In these specific scenarios, the traffic lights are controlled by services hosted in Raspberry Pis, which, in the real world, would either be hosted within the traffic light itself or in a separated equipment in the direct vicinity of the controlled traffic lights
- In both cases, a critical service is hosted on small, isolated spots of resources. By nature, those isolated resources could be cut off from the central clouds, either by accident or due to an attack
- This scenario precisely aims to demonstrate the ability of the proposed M&O architecture to efficiently handle cyber-security threats against a vulnerable application deployed at the extreme-edge



Demo #5 - Scenario 5.4:
MLOps techniques to deploy AI/ML service components

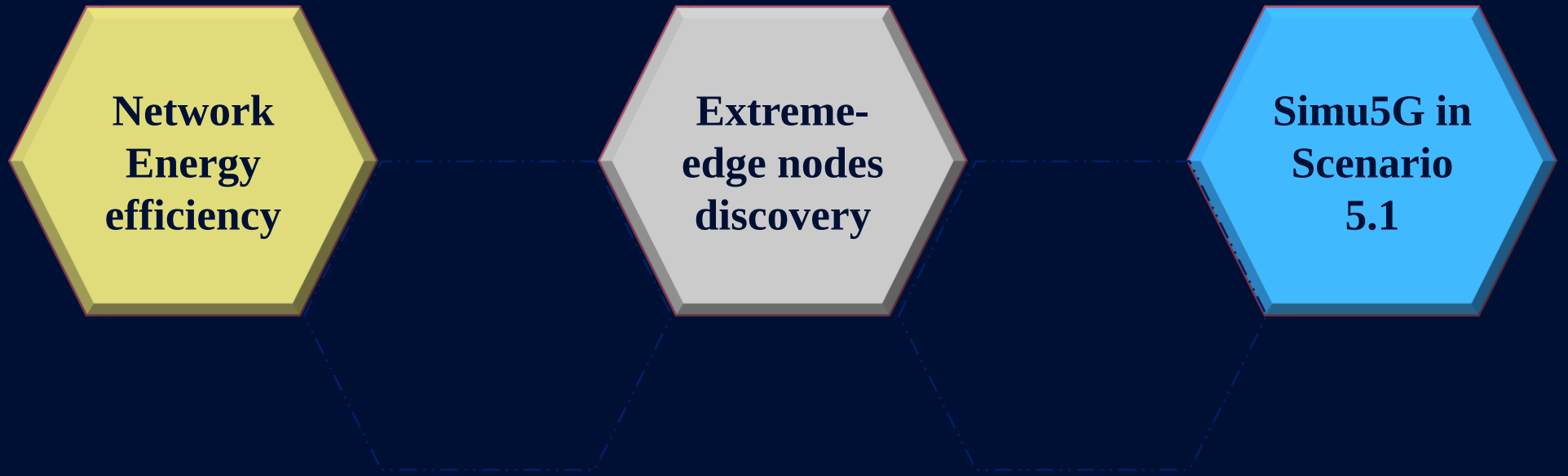
- This scenario targets the challenge of developing and deploying the AI/ML based services on the MNO premises using cloud-native DevOps techniques (MLOps), which can represent a challenge by itself in a typical multi-stakeholder telco-grade scenario
- One of the challenges this scenario targets is the workflows orchestration considering the main involved stakeholders (MNO and SW Vendor)
- This includes the processes for collecting and sharing the data sets needed to design and train the AI/ML models, as these processes are not considered in the regular MLOps workflows (which typically consider single stakeholder scenarios)
- The scenario targets a specific use case where a single SW Vendor develops, trains, and deploys a supervised AI/ML model on the MNO infrastructure, showcasing the whole MLOps cycle
- This scenario also covers a simple model drift managemet use case, which automatically redeploys the model when a drift situation is detected



Complementary Lab Experiments

Complementary Lab Experiments - Overview

- To complement the work adressed in Demos #4 and #5
- To explore one of the quantifiable targets assigned to WP6 in the Hexa-X work plan, namely “Improvements on the Network Energy Efficiency”
- To explore also some other topics that were considered interesting in the WP6 consortium, including the automated extreme-edge resources discovery mechanisms (closely related with the extreme-edge volatile resources orchestration), and the possible integration of the radio part in Scenario 5.1
- The following complementary experiments were conducted:



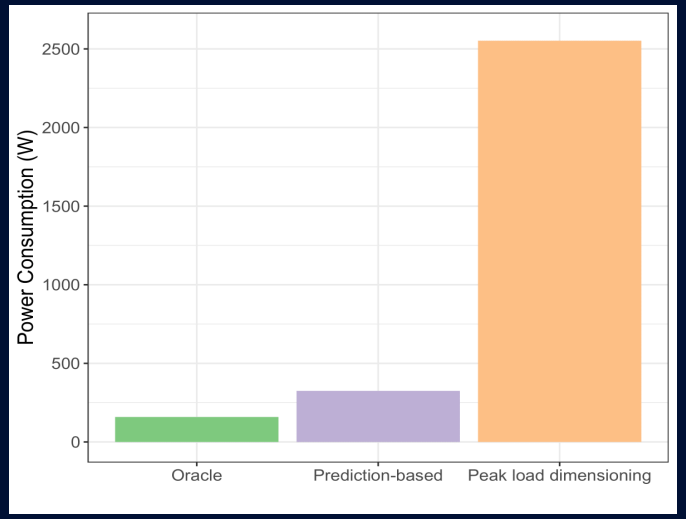
Lab Experiment 1
Network Energy Efficiency

This experiment consists of the simulation of a carrier-grade server's farm architecture with 150 servers, each one supporting up to 16 simultaneous requests. In this architecture, the following three scenarios were considered:

- Peak-load dimensioning (i.e., no orchestration). This is the worst-case scenario, where all servers consume the maximum energy. In this case, all servers are active regardless of the traffic demand. This causes a complete waste of resources even though the QoS is guaranteed.
- Oracle. In this case, the exact number of servers needed to match the QoS requirements are calculated. Therefore, the orchestrator accommodates the current traffic demand with a sufficient number of resources resulting in a more efficient approach. However, it is considered that this method is not realistic (or applicable in all cases) since knowing the current load can be challenging.
- Predictive orchestration. In this case, the traffic load is predicted based on the load history. This implementation is intended to be more realistic than the previous one. This approach is based on using an LSTM Recurrent Neural Network (RNN) to perform the load predictions.

Result:

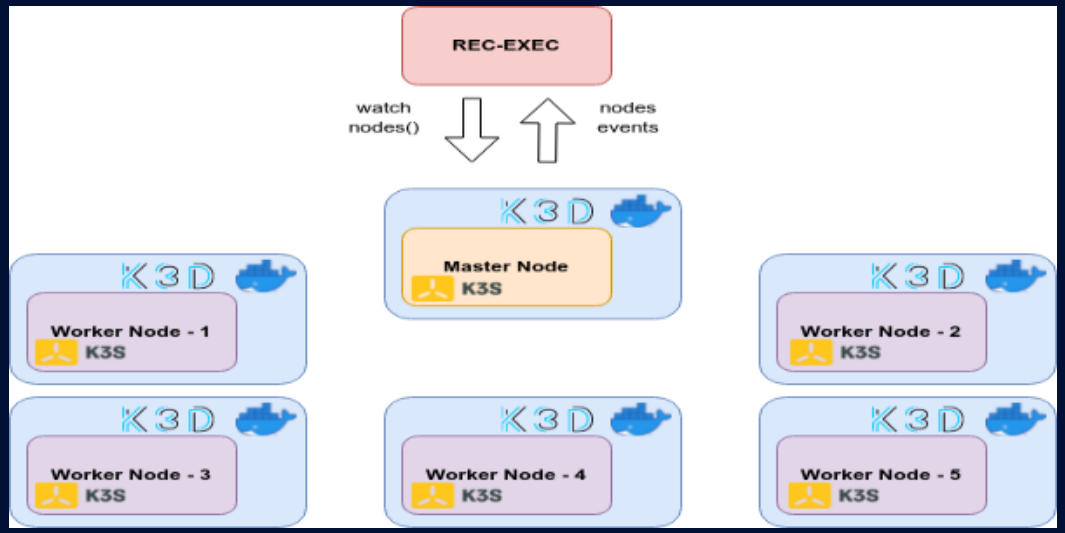
- For the **Peak Load dimensioning** scenario, the minimum number of resources to accommodate the peak-load demand is calculated
- For the **oracle** scenario, the exact number of resources for each traffic demand is calculated
- For the **predictive orchestration** scenario, the next traffic demand and scale-up resources are predicted
- This results in a 324.769 W of prediction-based solution getting very close to the optimal solution in the oracle scenario



Lab Experiment 2
Extreme-edge nodes discovery

- The discovery of extreme-edge nodes can have a role in the monitoring stage of a control loop.
- The information about nodes joining or leaving the infrastructure can be notified to components working at the network and/or service layers within the analysis and decision stages of the control loop, which can rely on AI/ML-based algorithms.
- The information related to the missing availability of nodes can be used at the upper-layer service orchestration to trigger migration actions, and, in case of poor accuracy, they may lead to unnecessary delays or breaks in the service execution and continuity.
- In order to test, validate and measure the performance of the node discovery feature, the K3S tool is used to emulate a scenario with a K8s cluster composed of a master node and five worker nodes, each representing volatile extreme-edge resources

Event	Average time	Max time	Std Deviation
Node joining the cluster	4.9s	6.7s	0.46
Node leaving the cluster	1.4s	2.7s	0.39



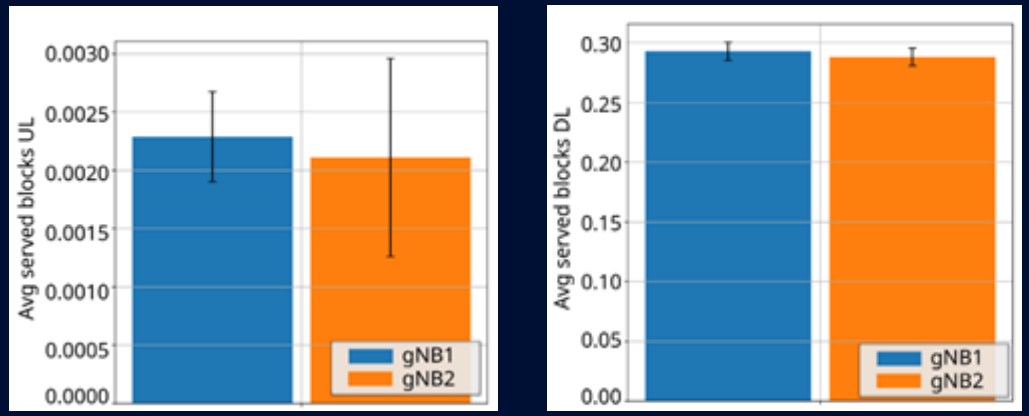
Main results:
It has been measured that the average time needed to spawn a node with K3d, including the time to discover the new node itself, is 4.9 s, with a maximum of 6.7 s and a standard deviation of 0,46. The leaving time is 1.4 seconds on average, with a maximum of 2.7s and a standard deviation of 0.39 (see Table 6 2). The additional overhead required to synchronize the resource inventory is negligible (in the order of milliseconds).

Lab Experiment 3
Simu5G in Scenario 5.1

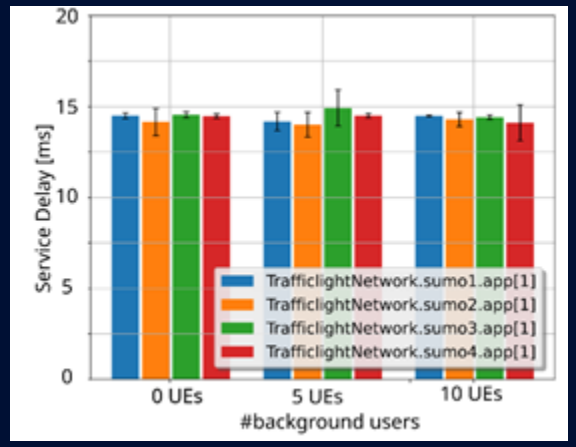
The evaluation of the impact that the B5G/6G RAN might have on Scenario 5.1 is of paramount importance in order to clarify future scenario enhancements aiming at achieving a higher TRL and full integration with the B5G/6G mobile networks stack, i.e., adding a real RAN to the scenario, implementing it on a real-life scenario, etc.

Experiment results:

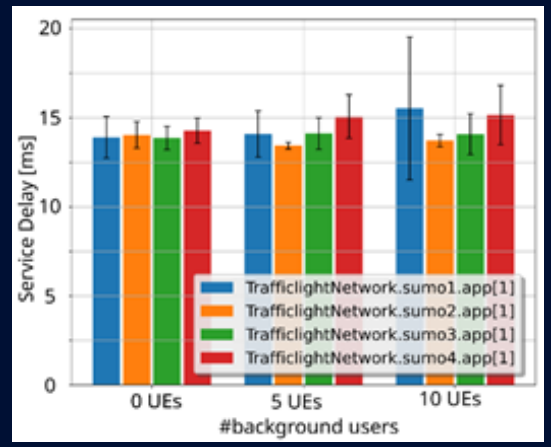
- The Simu5G impact on the network resources is very low
- The service delay is always below 16ms, even for the higher traffic loads, thus confirming the feasibility of the proposed methodology



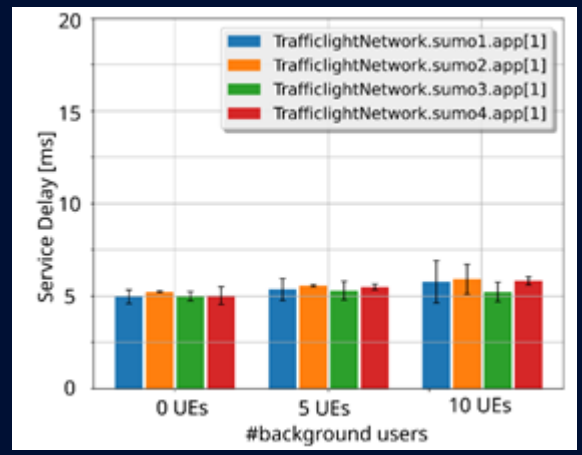
Average number of resource blocks in Uplink (left) and Downlink (right)



Average delay of the communication between the *SUMO Extreme-edge* components and the *RL Agent*



Average delay of the communication between the *Traffic Lights Control Logic* and the *SUMO Extreme Edge*

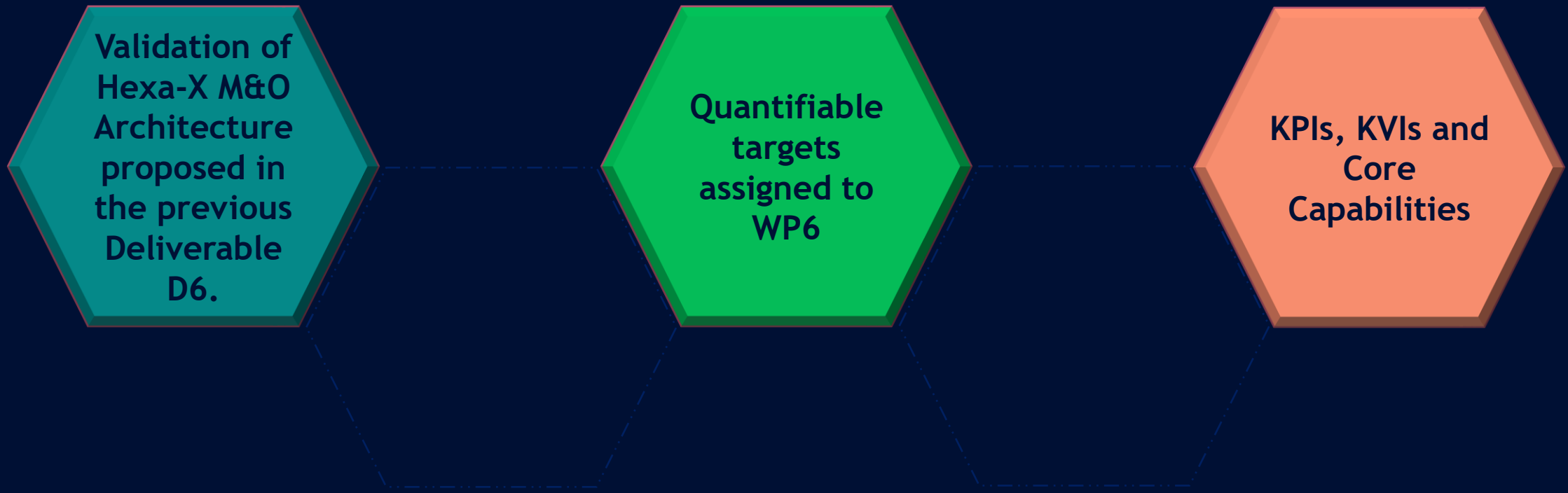


Average delay of the communication between the *RL Agent* and the *Traffic Lights Control Logic* components

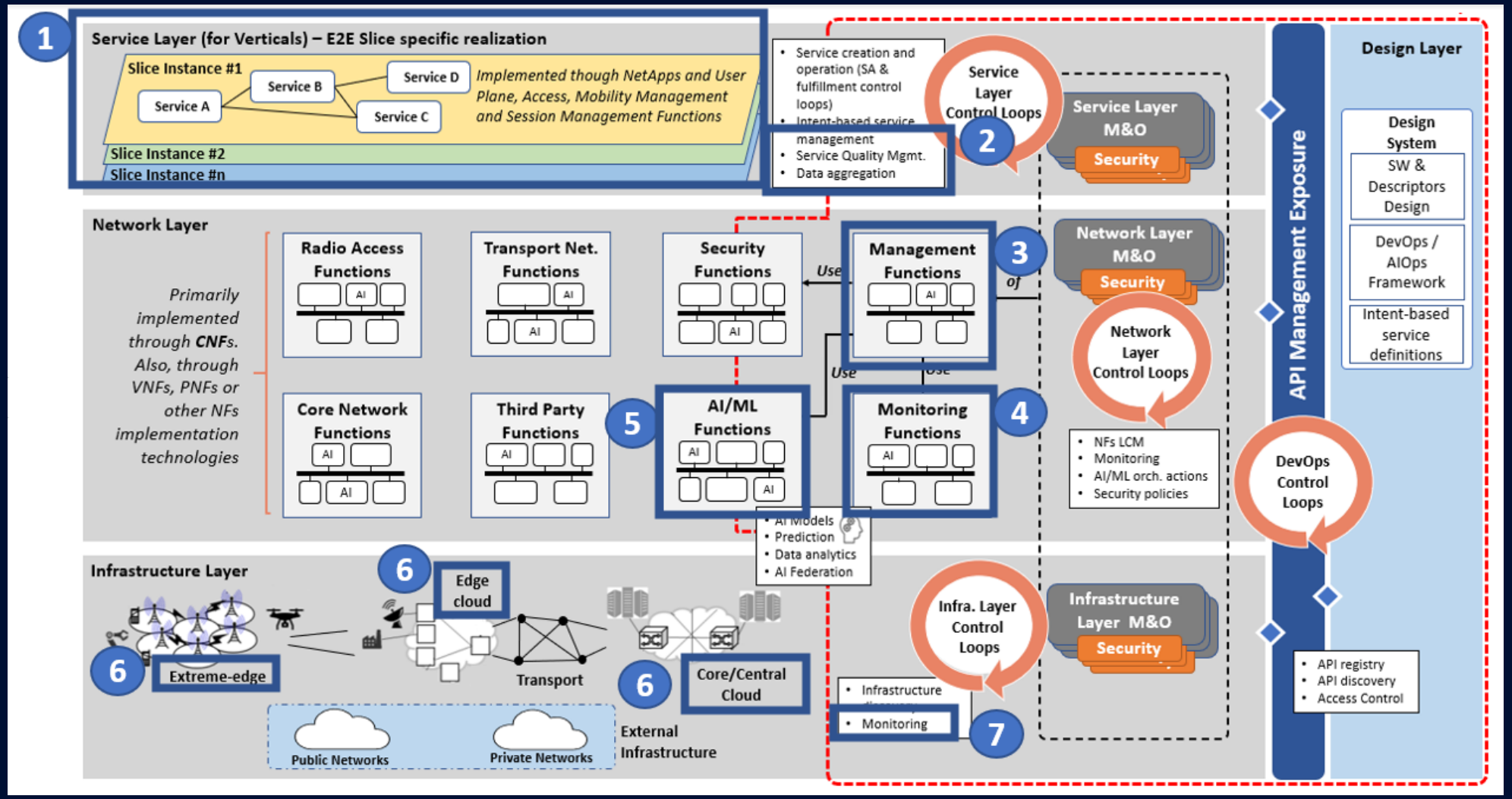
Evaluation

Evaluation - Overview

The evaluation has been performed based on the following:

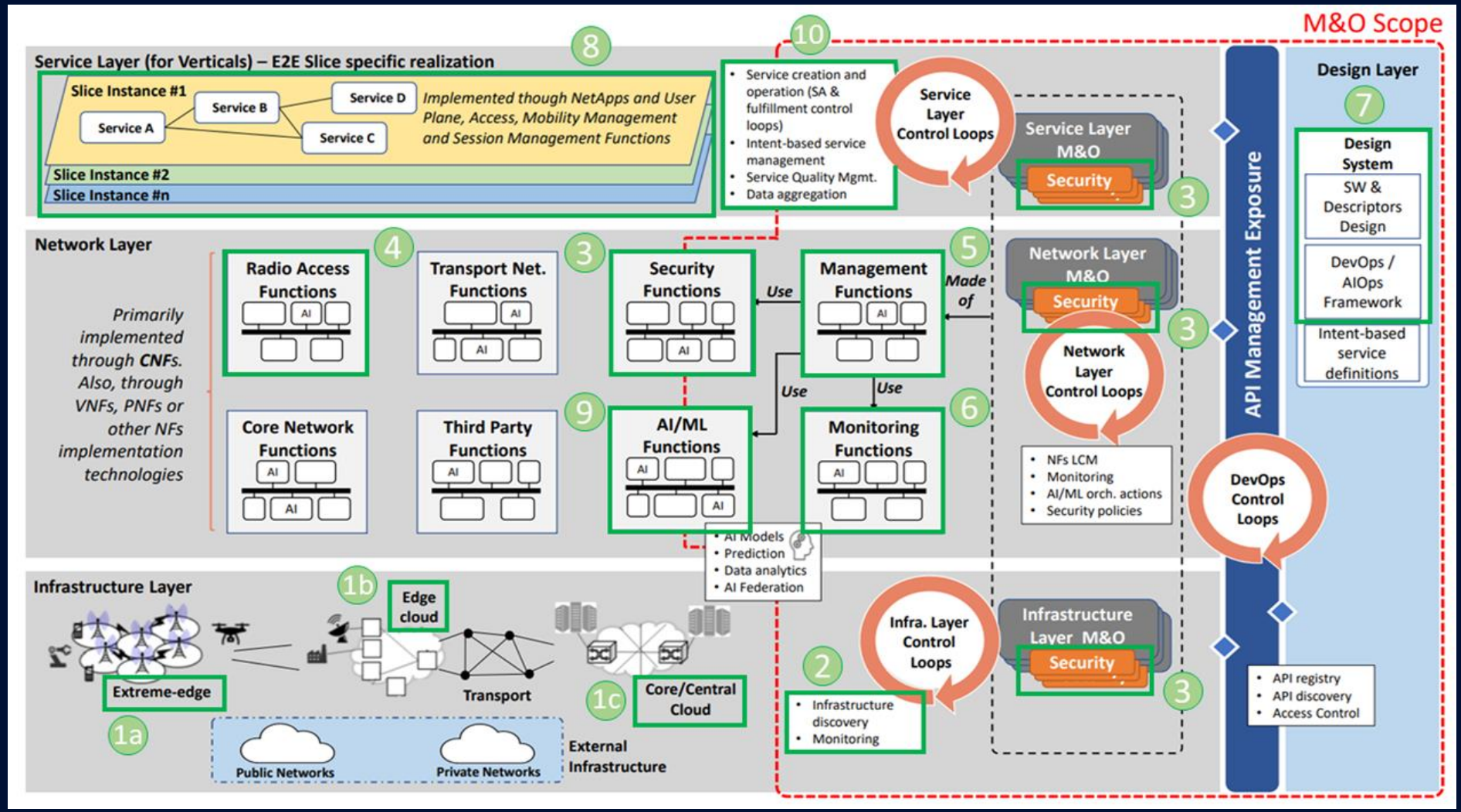


Validation of Hexa-X M&O Architecture - Demo #4



- It has been verified that Demo #4 can be implemented according the M&O architectural design in the previous Deliverable D6.2
- The figure shows the components in the architectural design implemented through this Demo (blue circles)

Validation of Hexa-X M&O Architecture - Demo #5



- It has been verified that Demo #5 can be also implemented according the M&O architectural design in the previous D6.2.
- The figure shows the components in the architectural design implemented through this Demo (green circles).

Quantifiable Targets Validation (1)

QT3a: Improvement in network reconfiguration times

- This QT requests to validate the feasibility of achieving network reconfiguration (regarding creation, composition, and scaling times) to be performed by (>10%) on the prediction horizon.
- Validated in Demo #4 with an additional performance improvement caused by the introduced components and architecture is the reduction of service downtime by more than >10%. Detailed results in D6.3 document.
- Also validated in Scenario 5.2. Detailed results in D6.3 document.

QT 3b. Improvement in time to onboard/remove resources from other domains

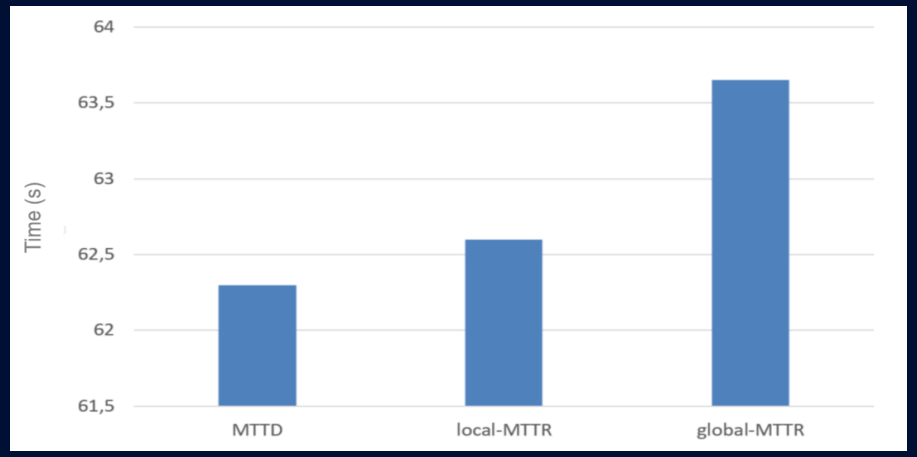
- This QT requests to validate the feasibility of improving by (>90%) in time the onboarding of new resources from other domains and managing the addition/removal of elements from the network.
- For this QT, the considered baseline is the legacy approach, i.e., the deployment of a service without using these cloud-native DevOps methodologies, which are compared with the results obtained by applying the MLOps techniques in the Scenario 5.4.
- This QT is considered achievable, however, further research should be done on more specific ways the DevOps methodologies could be applied in the telco-grade environment since the regular DevOps approach is typically applied “within” single organizations, promoting the joint work of their operational and development teams.

Quantifiable Targets Validation (2)

QT 3c: Improvement regarding service continuity

Scenario 5.3: Fast reaction response need is a primary concern for the proposed solution: while the central loop focuses on providing a long-lasting solution for the incident through eradication, in parallel, the local loop is dedicated to providing a temporary solution as fast as possible through mitigation. In order to measure the effectiveness of this system, we measured the relevant response times of our system:

- Mean Time To Detect (MTTD)
- Local Mean Time To Respond (local-MTTR)
- Global Mean Time To Respond (global-MTTR)



QT 3d: Improvement of the network energy efficiency using predictive orchestration

- This QT validates the feasibility of *increasing the network energy efficiency by (>50%) applying predictive orchestration* techniques
- Based on complementary lab experiment (Network Energy efficiency), it can be stated that by adapting to the forecasted demand, it is actually possible to reduce the power consumption to approx. 324.769 W, which represents approximately 88% of the original energy consumption, beyond the QT definition requires.

KPIs, KVIs and Core Capabilities

List of KPIs, KVIs and Core Capabilities adressed in the conducted demos and experiments

Experiment	KPIs	KVIs	Core Capabilities
Demo #4	• Programmability [%] • Processing capacity [s] • Creation Time [s] • Automation [Degree]	• Trustworthiness • Sustainability	• Integrated Intelligence • Use of embedded devices • Flexibility
Demo #5	• Programmability [%] • Processing capacity [s] • Creation Time [s] • Automation [Degree] • AI/ML models training time [s] • Maintainability [Degree]	• Trustworthiness • Sustainability • MTTR (Mean Time To Respond)	• Integrated Intelligence • Use of embedded devices • Flexibility
Network Energy Efficiency	• Latency [s] • Reliability [%] • Energy efficiency [W]	• Trustworthiness • Sustainability	• Integrated Intelligence • Flexibility
Extreme-edge nodes discovery	• Programmability [%] • Automation [Degree]	• Trustworthiness • Sustainability	• Integrated Intelligence
Simu5G in Scenario 5.1	• Latency [s]	• Trustworthiness	• Use of embedded devices • Flexibility

Future Work (1)

- More work on tools and technologies that could be used to implement the proposed architecture and the possible standards to align with
- Additional work on the API Management exposure
- Intent-based networking needs implementation and validation
- MLOps techniques for the deployment of other well-known ML paradigms (e.g., RL and FL). Exploring anonymization and encryption techniques, multi-vendor approaches MLOps approaches, targeting full zero-touch automation (ZTA)
- Research on extreme-edge nodes discovery
- Improving resource orchestration of extreme-edge nodes in end-to-end scenarios, e.g., considering the impact of their mobile connectivity on the management interactions between platform and worker nodes, enhancing the resource allocation logic with constraints related to per-node characterization, and using ML techniques to predict the time-variable attributes
- Coordination of automation and closed-loop decisions and actions across multiple domains, infrastructure layers and time scales and
- Coordination of management actions for data collection/transfer/storage and ML pipeline automation in distributed and multi-domain environments, taking into account data characteristics (ownership, sharing policies, privacy, etc.)

Future Work (2)

- Evaluation of a large-scale orchestration of different types of edge resources, having diverse resource capacity, performance, operation cost and availability, and providing services to heterogeneous user applications
- Using localization information, possibly linked to the node discovery functionality, to enhance the operations of resources and network functions placement
- Extending the reach of the performance diagnosis and functions placement mechanisms, shown in Scenario 4.2, from the services all the way to the network and the M&O components as well
- Further study on ‘network-of-networks’, coordinating the monitoring and control actions upon flexible topologies/networks, functions placement, and unified orchestration for the purpose of being able to provide ad-hoc networks to new orchestratable resources on-demand
- Further research, design and development in the two stages of the Level of Trust Assessment Function (LoTAF): network service selection based on user’s security and privacy requirements, intelligent optimization functions or technology-based threat analysis

Summary & Conclusions

Summary & Conclusions

- This report, as the final deliverable of the Hexa-X Work Package 6 (WP6), evaluates service management and orchestration (M&O) mechanisms for Hexa-X, described in Hexa-X D6.2
- This deliverable's main part consists of describing two demos (Demo #4, Demo #5) and other lab experiments
- The evaluation of proposed service management and orchestration mechanisms has been performed. The evaluation process includes the validation of the Hexa-X M&O architecture and discussion on achieved Quantifiable Targets, as well as related KPIs, KVIs and Core Capabilities
- Most of the novel capabilities in the M&O architectural design introduced in Deliverable D6.2 have been addressed (with the only exception of the application of the intent-based approaches for service planning and definition)
- The evaluations show that the assumed goals have been achieved. The report consists also of a section that describes the lessons learnt and suggestions concerning future work, as the report describes initial evaluations only, and still, many mechanisms and features of the proposed M&O framework have to be evaluated in detail

Thank you!

HEXA-X.EU

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 101015956.