An Optimisation as a Service Platform for Beyond 5G and 6G Networks

O. Sallent[(1)], J. Pérez-Romero[(1)], I. González[(2)], A. Santiago[(2)], J. Baliosian[(3)], L. Diez[(4)], R. Agüero[(4)], A. Muñiz[(5)], L. M. Contreras[(5)], E. Fernández[(6)], J. Moreno[(6)]
sallent@tsc.upc.edu, jordi.perez-romero@upc.edu, ivan.gonzalez@nemergent-solutions.com, adrian.santiago@nemergent-solutions.com, baliosian@gmail.com, ldiez@tlmat.unican.es, ramon@tlmat.unican.es, alejandro.muniz@telefonica.com, luismiguel.contrerasmurillo@telefonica.com, efernandez@e-lighthouse.com, jmoreno@e-lighthouse.com
[(1)] Universitat Politècnica de Catalunya, Barcelona, Spain. [(2)] Nemergent solutions, Bilbao, Spain.
[(3)] Universidad de la República, Uruguay. [(4)] Universidad de Cantabria, Santander, Spain.
[(5)] Telefónica Innovación Digital, Madrid, Spain. [(6)] E-lighthouse Network Solutions, Cartagena, Spain.

***Abstract*- The evolution towards 6G will require changes in the way how networks have to be operated to face the stringent service demands while at the same time using efficiently the available resources. In this context, this paper presents the Optimisation-as-a-Service (OaaS) platform being designed by the OPTIMAIX project that makes use of the Network Digital Twin (NDT) concept to test network configurations in a controlled environment. After describing the OaaS platform architecture, the paper presents the optimisation algorithms considered by the project and summarizes the use cases developed in the prototyping activities.**

I. Introduction

The continuous evolution of communications networks in the race towards 6G will introduce important operational challenges for network operators. Until recently, the main driver for networks design and operation has been the guarantee of throughput since bandwidth-based services have been dominant so far. With the advent of 5G, delay-based services have started to become relevant as well, as a consequence of the offerings of low latency and interactive services, such as tactile internet or augmented and virtual reality. Now, the advent of forthcoming beyond 5G (B5G) and 6G services is expected to introduce more stringent requirements motivating a new radical change on operations, adding new dimensions to service provision, such as delay-variation or precision-based services.

All this complexity triggers the need of introducing smarter, faster, and educated decisions on the operational processes, pursuing an overall optimisation on the usage of both networking and compute resources. In this context, the OPTIMAIX project [1] intends to develop an Optimisation-as-a-Service (OaaS) platform providing a set of tools for supporting the automated planning, operation and optimisation of 6G networks targeting the efficient use of available network and compute resources. The initial conception of this platform assumes a number of optimisation algorithms, possibly supported by Artificial Intelligence (AI) and Machine Learning (ML) tools. Each algorithm will address a certain planning or optimisation problem that, based on inputs such as the Service Level Agreement (SLA) requirements, the current load in different parts of the network, etc. will generate specific decisions regarding the configuration of the network that will be enforced through the corresponding management entities and interfaces.

The OaaS platform exploits the concept of Network Digital Twin (NDT), which is gaining traction as a perfect complement to the foreseen complex operation of future 6G networks. An NDT provides a virtual and updated representation of the network that brings the possibility to analyse, diagnose and emulate the physical network in a zero-risk environment [2]. For instance, the telco industry is now adopting the DevOps paradigm, supporting the Continuous Integration / Continuous Delivery (CI/CD) of services, introducing a dynamicity on the configuration actions realized over network and services never seen before. That high dynamicity of actions should be performed with extreme care for avoiding side effects and service affection on a critical infrastructure such as the telecom network. Then, an NDT can make available a replica of the network that can help to anticipate, validate and predict actions in the real physical network but in a controlled environment.

In this context, this paper presents in Section II the architecture of the OaaS platform as considered in OPTIMAIX project. Then, the paper outlines in Section III some examples of optimisation algorithms to illustrate the usefulness of the OaaS platform, while Section IV describes the on-going work to develop the prototype of this platform. Finally, Section V summarises the conclusions.

II. OaaS Platform Architecture

The OaaS platform, which is illustrated in Fig. 1, provides a novel and scalable solution for the optimisation of network services and resources including both the Radio Access Network (RAN) and the transport network. It is conceived as a hybrid system where one logically centralized entity (OaaS Master) and multiple federated ones (OaaS Nodes) coexist. The platform integrates computing clusters for the OaaS Master and OaaS Nodes, facilitating the deployment of Docker containers from remote repositories to execute optimisation algorithms and NDTs.

The OaaS Master includes the *OaaS platform Mgmt* module that consists of a Kernel that manages and reroutes all the requests within the OaaS platform. It includes two databases (DB). The *DesignDB* serves as a repository for network-related information, including data models of the SLA requirements and the elements of the network topology, amongst others. To optimize algorithms and NDT simulations, this database needs to have accurate network and infrastructure information in near real-time. In turn, the *Business Logic DB* manages and stores platform metadata, such as the locations where algorithms or NDTs are deployed and the specific paths they follow. This helps to ensure proper operation and efficient management of the platform.

The OaaS platform presents Application Programming Interfaces (APIs) for the integration with other modules through a RESTful architecture, using a human-readable format as JavaScript Object Notation (JSON). This approach

is extensively adopted in similar contexts due to its numerous benefits such as lightweight implementations, robustness, fast processing, and easiness of documentation and testing process. The current OaaS architecture includes two large blocks of APIs as depicted in Fig. 1.

Firstly, the *Resource management API* is found both in the OaaS Master and in the federated OaaS Nodes and it encompasses three key modules that allow the registration, instantiation and execution of resources. They are respectively the *Image management* module, the *Instance management* module, and the *Execution management* module. In turn, the *Integration API* involves three centralized modules in the OaaS Master. The *Design handling module* details the main operations to manage the network topology designs stored in the *DesignDB*. The algorithms and NDT executions will be based on these designs to recreate the real conditions over virtual environments. The *OaaS platform management module* enables the federation of OaaS Nodes within the OaaS platform, allowing the deployment and enrolment of new modules. Finally, the *South Bound Interface (SBI) Management module* allows registering the underlying components that manage the network resources and the computing resources of the network.

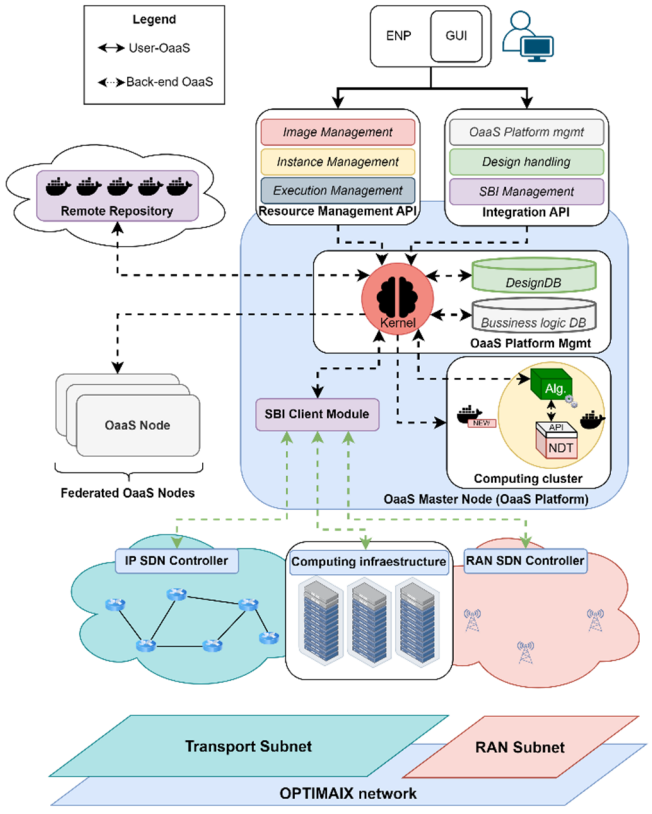

Fig. 1. OaaS platform schema and APIs

The acquisition of topological information from networks or the importation of existing computing resources is paramount to understand the resources present in underlying networks. To facilitate this, the OaaS platform records the IP addresses and communication ports where these resources are located, along with the access credentials and principal endpoints for executing requests. Then, the *SBI Client Module* enables sending requests to various active registered resources and abstracting the data from the received response. This necessitates a prior preprocessing phase, which allows for data optimisation before it is stored in the corresponding database.

Each OaaS Master or OaaS Node includes a *computing cluster* to accommodate the Docker instances with a variable number of resources such as algorithms and NDTs with AI/ML assisting tools. These resources are clustered in repositories and are managed as microservices instantiated via docker-compose. The algorithms address network problems such as dynamic resource allocation, network provisioning, capacity planning, etc., as described in Section III. They make use of different input parameters such as SLA, network slice components, network segments, current load or traffic demand in different parts of the network. AI/ML-based algorithms require a training stage to accurately adjust the existing models and enough network observation times to acquire all the needed data. During the training, these algorithms generate actions to be simulated in a NDT. Later on, the trained models can be applied into the real network components.

III. Optimisation Algorithm Examples

A. An algorithm for capacity sharing for RAN slicing

Network slicing is a key feature of 5G and beyond systems enabling the creation of multiple end-to-end logical networks, referred to as network slices, on top of the same physical infrastructure, each one optimized to the requirements of specific services and application domains. A network slice includes a core network subnet instance and a RAN subnet instance, denoted as RAN slice. The deployment of RAN slices has to deal with the management of the radio resources available in the existing cells in accordance with the requirements of each slice. For this purpose, capacity sharing mechanisms are needed to dynamically modify the amount of resources allocated to each RAN slice in each cell, ensuring an efficient use of the available radio resources and at the same time the fulfilment of the RAN slice requirements.

The algorithmic solution to the capacity sharing problem has been approached by means of Multi-Agent Reinforcement Learning (MARL). Among the existing Reinforcement Learning (RL) methods for deriving the policy of each agent the Deep Q Network (DQN) algorithm has been selected due to its ability to support high dimension state and action spaces. A detailed description of the algorithm is provided in [3].

B. An algorithm for addition of new cells in the RAN

The problem considered here assumes that, at a certain point of time and as part of the dimensioning process, the Mobile Network Operator (MNO) decides to expand the topology of the RAN by adding one or more cells. This can typically respond to the need to extend the coverage footprint of the RAN or to increase the deployed capacity.

Considering that, as discussed in the previous subsection, the MNO can have an ML-based capacity sharing optimisation algorithm, a challenging situation arises if the deployment of new cells changes the environment and this leads to a mismatch between the environment used for training the policy and the actual environment. Therefore, the previously learnt policy will no longer be optimum and can even lead to significant performance degradations. Consequently, a re-training process is needed to learn a new optimum decision-making policy. In order to efficiently re-train a deep RL-based capacity sharing policy when the MNO decides to expand the RAN with new cells, Transfer Learning (TL) techniques have been identified as key tools, since they allow accelerating the training process by re-using previous knowledge of a source task to perform a new target task. In this way, the re-training can start from the previously learnt policy and update it

according to the new environment. A detailed description of the algorithm is provided in [3].

C. An algorithm for dynamic functional split selection

In [4] we jointly consider network slicing and functional split selection. The proposed solution aims at increasing the centralization degree, by moving more functions from the radio unit (RU) to the distributed unit (DU) as well as to the centralized unit (CU). This would enable a tighter cooperation between RU, yielding potential benefits, for instance reducing the potential interference. In this sense, we took the reduction factor reported in [5], so that higher centralization degrees might bring a remarkable gain in terms of interference. On the other hand, the available computation resources at those nodes (DU/CU) need to be also considered, since higher centralization would actually require more computing capacity. The proposed algorithm, which could be easily integrated within the OPTIMAIX OaaS, aims at exploiting this disaggregation paradigm to guarantee that the heterogeneous requirements of different slices can be fulfilled. We introduce a dynamic solution that exploits the Lyapunov's Theory, suitable for uncontrolled random environments. The implemented approach, which uses the GLPK library to solve the corresponding optimization problems, was compared in [4] with static alternative solutions, based on the most widespread functional split configurations. A heterogeneous network, with various overlapping areas was used to assess this performance comparison. The obtained results evince that jointly considering network slicing and functional split configuration guarantees the fulfilment of the performance requirements, while strongly reducing the required radio resources, i.e. the number of physical resource blocks (PRB), which would bring energy savings. This result is observed for different cluster capacities, and for two values of the maximum spectral efficiency. As mentioned above the proposed technique exploits well known optimization techniques and it could thus be integrated within the OPTIMAIX framework, as a network operation module, which requires a near-real-time feedback to be sent to the real network.

D. An algorithm for end-user services placement

The algorithm introduced in our prior work [6] tackles a crucial challenge related to end-user services. Each service comprises a collection of microservices designed for deployment on cloud-network slices [7]. These microservices are distinguished by four essential technical dimensions: storage, representing the storage requirements of each microservice; bandwidth, reflecting the data generation or consumption rate; computing capacity necessary for execution; and maximum permissible delays between pairs of microservices. Considering the diverse demands of individual microservices within an end-user service, it becomes imperative to deploy them on distinct infrastructure elements, each supporting a specific cloud-network slice. Our research operates under the assumption that the telecommunications operator's objective is to streamline the distribution of workload within the network. This optimization aims to enhance overall network efficiency during service provisioning, ensuring a satisfactory user experience, and ultimately maximizing profit. In envisioning a network with various Points of Presence (PoPs), our approach involves distributing service components across different network locations. To address this service allocation problem, we devised a strategic solution by extending our approach to a related yet simplified problem. This extension, termed Vector Successive Shortest Path (VSSP), is a heuristic derived from the well-established Successive Shortest Path (SSP) algorithm. Our simulations validate the efficacy of our proposal, demonstrating that our approach yields operator profits close to optimal. Notably, the computational efficiency of our solution, as obtained through VSSP, is orders of magnitude faster than traditional methods. This innovation in service allocation holds promise for telecommunications operators seeking a balance between optimal profit and efficient resource utilization.

E. An algorithm for exposure of time-variant routing information to applications

The algorithm here described is based on the work presented in [8] as a standardization proposal for anticipation in topology changes for Time-Variant routing scenarios.

In dynamic network environments such as satellite networks and mobile communication networks, where routes are subject to periodic variations, selecting network routes accurately becomes challenging. Therefore, utilizing AI based engines in the prediction of future network behaviour could help to anticipate potential service limitations (i.e., increase in latency, decrease in network or compute capabilities or connection loss) and migrate services before service drops. For that reason, we propose the integration of Application-Layer Traffic Optimization (ALTO) as a topology exposure technology and an NDT as the technology to predict potential resources issues or underperformance.

This algorithm addresses two scenarios: resource request instantiation and network changes prevention. In both cases, a Network Controller collaborates with ALTO, NDT, and network elements to predict and take actions: while NDT provides load and topology predictions, ALTO generates Cost and Network Calendars based on network topology and predicted changes. In parallel, the current view will be monitored using network connectivity protocols (e.g., BGP, LLDP, etc.) to show the current real-time topology and therefore provide a more completed view to the client, mixing both real state information and forecasted metrics.

IV. Next Step: Prototyping

OPTIMAIX project started in January 2022 and will end in December 2024. After defining use cases, architecture and a bunch of solution algorithms, the focus of the project during 2024 is placed in prototyping. A network planning-oriented and a network operation-oriented use cases will be showcased in the implemented OaaS platform. Details of the integration process and tests to conduct are provided in [9]. In the following, a description of the use cases to be implemented is provided.

A. Use case #1: network planning

This use case has been designed to validate and show the functional capabilities of the OaaS platform for a network planning scenario. The main purpose is to demonstrate that the building blocks of the OaaS platform have been successfully integrated, validating that the OaaS platform can successfully perform its functional duties for network planning optimization.

The problem targeted by this use case consists of the optimal deployment of all the microservices that compose an end-user service on top of the infrastructure of a network provider, as illustrated in Fig. 2. Each of these microservices is characterized by a set of computational and networking constraints that must be satisfied for the correct operation of

the service. On its side, the telco provider infrastructure made available for deployment consists of computational centres and communication links distributed across the network edge to the cloud. In that framework, the target is to find the appropriate computational and networking resources in the substrate infrastructure to optimize the allocation of the set of microservices fulfilling their constraints, while minimizing the cost derived from the allocation.

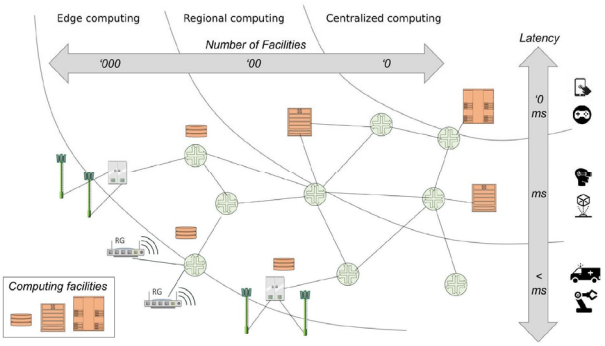

Fig. 2. Potential service placement options

In the specific use case envisioned, we adopted a Mission Critical Service (MCX) as the end-user service to be deployed and the end-user services placement algorithm described in Section III D, as the core to confer the optimization capabilities to the OaaS platform. The network planning problem addressed will allow the deployment of the MCX service attending to its specific necessities and constraints, guaranteeing its proper performance, while making minimum use of the network provider's infrastructure. This specific use case shows a double benefit through its optimization, namely the reduction in cost for the network provider by avoiding to waste more resources than truly needed, but also the certainty for the end-user service that the deployment will comply with its needs and constraints in terms of computation and networking.

B. Use case #2: network operation

This use case introduces the application of an NDT within an Operation and Management (O&M) framework for telecom operators, like in Network Operations Centres (NOCs), focusing on predictive network status assessment to aid in anticipating time-variant related network behaviours. It is related with the algorithm described in Section III.E.

This approach aims to enhance integration between applications and networks, providing valuable insights by anticipating the consequences of expected changes for service providers. This network behaviour is known as Time Variant Routing (TVR), a concept addressing scheduled changes in network infrastructure, such as planned maintenance procedures or upgrades or periodical network topology modifications. The NDT facilitates predicting the network's post-change status, enabling proactive decision-making for optimizing service delivery.

The specific requirements for this use case involve the integration of the NDT within an OaaS platform, alongside, a Network Operation Application, a Change Scheduler and an ALTO server as shown in Fig. 3.

In this architecture, the NDT receives inputs including current network topology and planned changes, and outputs predictions of resulting routing topology and network metrics. The Change Scheduler interacts with the NDT instantiated in the OaaS platform, via the NDT API, to retrieve scheduled changes data and communicates with the ALTO server to facilitate data provision to the NDT. Finally, the ALTO server will provide both current routing topology information and the forecasted network status to the vertical applications. This use case enables topology changes exposure to facilitate the graphical management in operator's users.

The correct execution of this use case will carry the improvement of the service allocation, improving the decision-making from long-time services depending on future network status. This also can be translated into a reduction in the times a service is needed to be migrated and in avoiding evitable service drops.

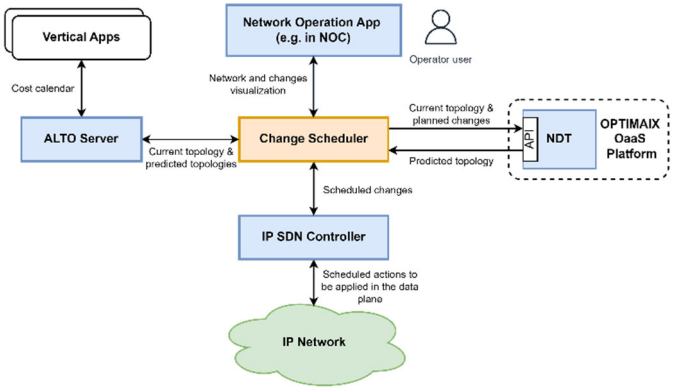

Fig. 3. Infrastructure proposed for the Network Operation Use Case.

V. Conclusions

This paper has presented the architecture of the OaaS platform developed by the OPTIMAIX project to support the automated planning, operation and optimisation of 6G networks. Through properly defined APIs and databases, the OaaS enables the execution of different optimisation algorithms and of NDT modules that allow analysing, diagnosing and emulating the physical network in a safe environment. In this respect the paper has also described the algorithms developed by the project as well as the use cases considered for prototyping.

Acknowledgements

This work is funded by the Spanish Ministry of Economic Affairs and Digital Transformation and the European Union - NextGenerationEU under projects OPTIMAIX_OaaS (Ref. TSI-063000-2021-34) and OPTIMAIX_NDT (Ref. TSI-063000-2021-35).

References

[1] OPTIMAIX project, https://optimaix.upc.edu/

[2] C. Zhou et al. "Digital Twin Network: Concepts and Reference Architecture", *Internet Engineering Task Force*, April, 2023.

[3] OPTIMAIX_OaaS Deliverable 2.1v1, Documento de diseño de las APIs y los algoritmos de la Plataforma, September, 2023.

[4] N. Villegas, S. Pérez, L. Diez, R. Agüero. "Joint and Dynamic optimization of functional split selection and slice configuration in vRAN". IEEE Wireless Communications and Networking Conference, WCNC'24

[5] A. M. Alba, S. Janardhanan, W. Kellerer, "Enabling Dynamically Centralized RAN Architectures in 5G and Beyond," IEEE Transactions on Network and Service Management, vol. 18, no. 3, pp. 3509–3526, 2021.

[6] J. Baliosian, L. M. Contreras, P. Martínez-Julia, J. Serrat, "An Efficient Algorithm for Fast Service Edge Selection in Cloud-Based Telco Networks," IEEE Comm. Magazine, vol. 59, no. 10, pp. 34-40, October 2021

[7] A. Garcia-Saavedra, J. X. Salvat, X. Li, X. Costa-Perez, "WizHaul: On the Centralization Degree of Cloud RAN Next Generation Fronthaul," in IEEE Trans. on Mob. Comput., vol. 17, no. 10, pp. 2452-2466, Oct. 2018.

[8] L. M. Contreras, "Using ALTO for exposing Time-Variant Routing information", Internet Engineering Task Force, Oct. 2023.

[9] OPTIMAIX_(OaaS+NDT) Deliverable 2.2, Documento de especificación del proceso de integración y de los test a efectuar, December, 2023.