CLOUD FOR HOLOGRAPHY AND AUGMENTED REALITY

Antonios Makris[1], Abderrahmane Boudi[2], Massimo Coppola[3], Luís Cordeiro[4], Massimiliano Corsini[3], Patrizio Dazzi[3], Ferran Diego Andilla,[5], Yago González Rozas[6], Manos Kamarianakis[7], Maria Pateraki[7], Thu Le Pham[8], Antonis Protopsaltis[7], Aravindh Raman[5], Alessandro Romussi[9], Luís Rosa[4], Elena Spatafora[9], Tarik Taleb[2], Theodoros Theodoropoulos[1], Konstantinos Tserpes[1], Enrico Zschau[10], and Uwe Herzog[11]

[1]Department of Informatics and Telematics, Harokopio University of Athens, Greece
[2]Aalto University, Espoo, Finland
[3]CNR, Pisa, Italy
[4]OneSource, Coimbra, Portugal
[5]Telefonica Research, Barcelona, Spain
[6]Tecnologías Plexus, Santiago de C., Spain
[7]ORamaVR, Heraklion, Greece
[8]Collins Aerospace, Cork, Ireland
[9]HPE, Cernusco sul Naviglio, Italy
[10]SeeReal Technologies, Dresden, Germany
[11]Eurescom, Heidelberg, Germany

ABSTRACT

The paper introduces the CHARITY framework, a novel framework which aspires to leverage the benefits of intelligent, network continuum autonomous orchestration of cloud, edge, and network resources, to create a symbiotic relationship between low and high latency infrastructures. These infrastructures will facilitate the needs of emerging applications such as holographic events, virtual reality training, and mixed reality entertainment. The framework relies on different enablers and technologies related to cloud and edge for offering a suitable environment in order to deliver the promise of ubiquitous computing to the NextGen application clients. The paper discusses the main pillars that support the CHARITY vision, and provide a description of the planned use cases that are planned to demonstrate CHARITY capabilities.

Keywords Cloud computing · Edge Computing · Cloud Services · Cloud Infrastructures · Network technologies · Internetworking

1 Introduction and main CHARITY concepts

While electronic means of communications have become a commodity in modern societies, their availability and those of various Internet-based tools have proven to be a vital element in the current pandemic situation, to maintain social interactions, to keep businesses going, etc. Equipped with that experience and with the pandemic threat being an ongoing concern, work on novel technologies that will make communications more immersive has received a push. Such technologies must enable immersive communication applications like Virtual Reality (VR), Augmented Reality (AR) or Holography to become largely available, reliable, and commercially sustainable. From a network standpoint, they define a new class of services where best-effort and simple traffic differentiation approaches are insufficient to meet their strict requirements. Studies showed that for an acceptable user experience with high fidelity, the latency should be less than 15ms and the bandwidth ranges from 1Gbps up to 30Gbps. Even when considering the advances of 5G, these applications pose a huge challenge to the network and the entire computation infrastructure. One option to meet such demanding requirements is through the use of edge computing capabilities. Edge computing allows the processing of the data to take place closer to where the services are consumed and/or where the data are generated, thus reducing the bandwidth between data centres and sensors. To accelerate adoption and reap the benefits of edge computing, technologies from various domains need to be exploited to eventually realize the cloud/edge integration. These

include computing, network and application-oriented technology fragments that need to be ingeniously put together to form and manage a network and computing continuum. In this context, the EC-funded R&I project CHARITY has set off in order to address these challenges and to develop and prototype a number of related use cases. CHARITY aspires to leverage the benefits of intelligent, autonomous orchestration of cloud, edge, and network resources, to create a symbiotic relationship between low and high latency infrastructures that will facilitate the needs of emerging applications. Looking at the overall concept, the overarching vision of CHARITY is the development of a unified framework ensuring a complete cycle of highly interactive services management, spanning from CI/CD to life cycle management (LCM) and orchestration. The CHARITY ecosystem is defined as the heart of three main pillars:

1. End-user equipment and highly interactive and collaborative services and applications, where CHARITY will focus on designing, developing, and managing highly interactive services, enabling next generation applications by fulfilling their high demanding requirements.
2. Cloud/edge infrastructure and technology will be designed to achieve, among others, cost savings, increase service elasticity, and reduce software / hardware dependencies. Both Artificial Intelligence (AI) techniques and Zero-touch network and slice life-cycle management (ZSM) concepts will play a crucial role.
3. Optimising the Telecommunications infrastructure, to ensure the achievement of targeted KPIs and requirements of AR, VR and Holography enabled applications, e.g., through the use of classes of network traffic and prioritised flows.

2 Use cases and main challenges

Below, the use case (UC) applications, upon which the CHARITY platform will be tested, are described and their main challenges are identified. The UCs are organised in three main categories, namely Real time Holographic (UC1.1-3), Immersive Virtual Training (UC2.1-2) and Mixed Reality Interactive applications (UC3.1-2), aiming to address the main challenges in these sectors, ultimately allowing the validation, demonstration and showcasing of the edge/cloud computing innovations of the CHARITY project.

The *Holographic Concerts* UC1.1 utilizes a pseudo holographic projection system, based on the Pepper's Ghost principle [1], that creates an illusion of musicians playing live on a stage. Different band members (musicians) are situated in different locations, and are virtually taking part in a concert. Each musician's "2D hologram" is captured via a video camera and transmitted together with the sound stream, further processed in the CHARITY cloud, synchronized and projected to a dedicated display on stage. The video from viewers, situated in front of the stage is recorded and transmitted back to the band members as feedback. The UC is based on cloud and local processing supported by CHARITY services and exploits relevant software and hardware resources supported by the platform. Some modules and services may run remotely on the CHARITY platform, e.g. video mixing and composition, compression and rendering, while other locally on PCs at stage or at musicians location e.g. the synchronization service. The main challenge of this UC is to synchronize all video streams prior to displaying on the stage. Especially the audio synchronization is of importance, so separating audio from video data and handling with more priority is one of the options to implement. The goal is to achieve a flawless audio experience also in case of sudden bandwidth limitations, sacrificing video quality for audio.

The *Holographic Meetings* UC1.2 enables the main participant, acting as speaker, to be situated at any location and transmit its video and audio to numerous displays in various venues concurrently. In this case, the 2D video of the speaker is transmitted to a remote CHARITY service that transforms and renders it accordingly, for any type of holographic display connected to the CHARITY platform. The local PC at speaker's location supports reception of unedited video feeds from audience locations, to enable visual communication between the Speaker and the Audience in real-time. This UC is a simplified version of UC1.1, where no synchronization is needed, but the main challenge is to be able to support multiple output display devices concurrently.

In the *Holographic Assistant* UC1.3, a three-dimensional (3D) avatar is presented on a holographic 3D display (H3D) [2], that provides natural language replies and compiled visual 3D information, to user-spoken queries, after fetching results from 3rd party internet data services. Information and services offered by the holographic assistant are accessed through cloud-based services via APIs provided by CHARITY platform or 3rd party services available on the internet (e.g. weather, stocks or a chatbot). The H3D display is based on 3D Holography, that uses interference of light technology [3] to modulate coherent light and generate realistic visual representations of millions of 3D points in space, thus provides real depth. The visualized 3D data is extracted from a 3D point cloud stream received from CHARITY services. The use case, that is merely based on local side software, PC-hardware and H3D display (incl. Eye-Tracking), is supported by CHARITY cloud services that hosts assistant logic, Unity3D rendering, speech recognition, and 3D point cloud processing. The streamed 3D point cloud is received and directly displayed, as 3D hologram, on the H3D

display in real-time [4]. The main challenge of this UC is the real-time generation, compression/decompression of 3D point cloud data.

In the *VR Medical Training* UC2.1, multiple players execute predefined surgical scenarios in a VR environment, towards an enhanced [5] medical training experience [6, 7]. The UC's pipeline exploits CHARITY resources for advanced CPU and GPU processing for physics, rendering, compression supporting low latency, and increased bandwidth, specially targeting untethered HMDs with limited resources, GPU, battery and mobility. The application instance, deployed through two stateful micro-services on the edge-cloud - the Geometric Algebra inTerpolation Engine (GATE) and the Physics Engine - is responsible for computing, rendering, and encoding the images that will be transmitted to the HMD by a signaling server. In addition, run-time adaptation and dynamic optimization of the GATE is exploited based on the network characteristics [8]. The lightweight HMD is responsible for decoding and projecting the transferred images from the edge-cloud, and for capturing and transferring user events (e.g., controllers' position, triggers) to the application instance. The main challenge of this UC is the dissection of the Physics Engine from the Unity3D pipeline into a separate service that will enable lower running times.

Using the *VR Tour Creator* UC2.2, the user can build interactive virtual tour experiences and live streaming scenes in Virtual Reality. The application can be used for multiple purposes including learning, storytelling, marketing and real estate. The virtual tour supports 360° videos, panoramas, 3D models, standard images and videos, as well as basic 3D meshes. The application involves several modules both at the back-end and the front-end. The front-end modules consist of a web-app that manages and processes real time editing of the 360° videos tours created by the user, and a viewer app that enables the user to view and consume the content created by the web-app. The back-end consists of several components that are responsible for hosting media content, image processing, 3D model rendering, video format conversion and video streaming to the back-office. These components are deployed as containerized micro-services in CHARITY. The main challenge of this UC is to support faster video conversion and streaming through CHARITY platform's advanced processing capabilities, low latency, and increased bandwidth. Additionally, depending on the type of viewing device or the network characteristics, the 3D engine service needs to adapt its processing and rendering processes to different resolutions in order to serve multiple levels of mesh details.

The *Collaborative Gaming* UC3.1 provides a highly immersive multiplayer AR game. To provide players with sufficient immersion, a dedicated multiplayer engine is developed for synchronizing all dynamic game objects along with user's states throughout end devices. The overall solution requires the infrastructure to provide key features: very low network latency and efficient resource discovery service, a trusted infrastructure (cloud/edge) to support Game Server from dishonest player's breaches. UC3.1 explored the 3D Point Cloud technology to enrich gameplay and strengthen player's immersion. The built-in device cameras' capabilities are exploited to provide input and use the output data to mix the real and virtual environments. The Mesh Collider Generator Service supported by CHARITY is used to enable precise reconstruction of the real environment geometry within one game session based. The main challenge of this UC is to optimize the multiplayer engine by minimizing the amount of data sent over the network to maintain low round trip time among multiple players.

The *Manned-Unmanned Operations Trainer* UC3.2 enables a collaborative immersive training environment of emerging civil manned-unmanned teaming concepts while minimizing the involvement of expensive equipment. The key goal is to advance the deployment of training simulators on the cloud-edge continuum and target XR devices to deliver compelling immersive environments with minimal local technology assets. Trainees can virtually collaborate in a largely synthetic environment to perform coordinated search. They can also control remotely unmanned vehicles, such as simple aerial rescue drones, to approach inaccessible terrains in order to gain the situation awareness in the search-and-rescue scenario. The fluidity of CHARITY cloud/edge resources and network orchestration is leveraged to facilitate engagement and collaboration between multiple simulation instances. The main challenge in this UC is to virtualize the existing local simulation pipeline that will enable scalable collaborative training simulations.

3 System Architecture

Figure 1 depicts the general architecture of the CHARITY framework. It is composed of four main layers supported by an Application Management Framework (AMF) and a CI/CD pipeline.

The infrastructure layer, at the bottom, consists of the physical elements involved in the XR service, which spans from end user devices to the network transporting data, to the edge and cloud that are offering the computation. Due to the bandwidth-intensive and latency-sensitive nature of XR services, CHARITY aims at leveraging many cloud providers at the same time as a way for enabling edge-cloud continuum while also providing very high bandwidth, reliable and deterministic networking.

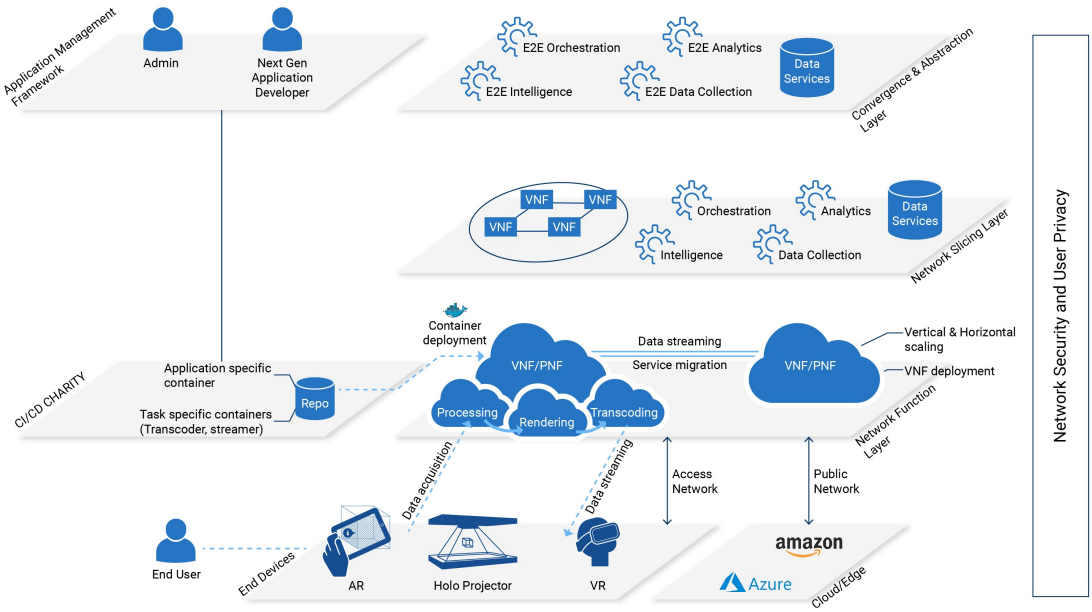

Figure 1: CHARITY's High-Level Architecture.

The second layer, named Network Function Layer (NFL), is responsible for abstracting the heterogeneity of the underlying infrastructure and thus, implementing the edge-cloud continuum concept by defining an orchestration framework that can seamlessly and efficiently run XR services. NFL provides a cloud native platform where XR services are implemented and ran as micro-services which permits great flexibility and efficiency. Indeed, a service can easily be relocated to match its requirements in terms of KPIs; also, an XR service can be split so a part of it is ran in the edge to reduce latency and bandwidth usage, while the compute-heavy load part of it is ran at distant clouds.

Given the fact that an XR service can be split across many domains, the Network Slicing Layer (NSL) is responsible to gather all of these domains and stitch them together to provide an end-to-end service. The first service automation loop is provided by NSL; it is present in each domain composing the XR service. It is implemented by following the OODA (Observe, Orient, Decide, Act) concept, which consists in data collection, analytics, intelligence, and orchestration. This local automation loop follows the ZSM [9] concept and thus, it implements self-orchestration, self-optimization and self-healing which enforces the satisfaction of KPIs for each sub-slice.

The Convergence and Abstraction Layer (CAL) is the E2E manager and orchestrator of XR services. It is responsible for enforcing the KPIs at the whole XR service level. This means that decisions of the composition of sub-slice, such as which sub-slice to use or where to run the sub-slice, are taken at this layer. For instance, having a latency budget, this layer will split that budget over the appropriate domains so the E2E latency is below the budget. Following the ZSM framework and in order to enforce the KPIs, CAL also implements an E2E automation loop. This loop's main objective is to enforce the E2E KPIs and to perform proactive actions in order to preserve the XR service from degradations. It also offers an interface from which XR developers and providers can submit their blueprints. These consist in description files that specify all the building blocks and their interconnection to form an XR service.

The Network Security and User Privacy Layer (NSUP) is used to secure the platform and the XR services running on top of it. This consists in securing the images of the software composing the XR service and the communication between its components; the protection of privacy of end users by processing their data in the edge and thus avoid sensitive data to be transported across all the network to be processed in the cloud. NSUP can also provide dynamic security where security components are dynamically added to the XR service.

Finally, the AMF and CI/CD pipeline, provided by CHARITY, is the entry point of XR providers and developers to the CHARITY platform. Among other, the AMF is used to define the blueprint of XR services and to launch, stop, modify and configure running XR services. While the CI/CD pipeline is used to ensure the deployment of a new XR service and/or a new version of a running service would not cause service degradation for that XR service or any other concurrent XR service. More details about the CHARITY architecture can be found here [10].

4 Edge and cloud infrastructure orchestration

Containerized components and microservices are largely promoted as the appropriate solution to efficiently deploy and manage novel applications on top of a hybrid edge/cloud infrastructure. This will most likely lead to a multi-component cloud application model where the components may be managed by heterogeneous host environments.

The primary need that emerges in such a system is for the orchestration of the deployment of its various components so that all inter-component dependencies are satisfied. After the deployment, monitoring and recovery mechanisms must be enacted to deal with the possible failures of the deployed components.

Previous experiments have reported various scenarios with applications deployed across different IaaS and PaaS providers and show that almost tweny percent of the scenarios experienced some failure. Other works have shown that the higher the number of components forming an application, the higher is the probability of its deployment to fail, due to the failure of one of its components.

The current support for operating cloud applications and recovering them from failures is however not fully automated yet. Usually workloads are restarted manually, perhaps in a new machine based on configurations in tools like Chef or Puppet, something that requires a considerable amount of setup and manual intervention. Automated solutions offered by Amazon, Google or Azure may do well in timely detecting the failure, but the mitigation measures basically consist in stopping the instance and creating a new one. These procedures are limited to components deployed using specific solutions, and no dependencies other than associated storing volumes are considered.

For the CHARITY project, a 5-way approach is planned:

4.1 Service placement

For which an Artificial Intelligence based Resource aware Orchestration (AIRO) framework is needed. The AIRO framework leverages the ZSM concept, cloud-native approach, and Machine Learning (ML) techniques for efficiently managing network and computation resources. This AIRO framework is implemented by deploying a monitoring and management agent alongside the master node at each cluster for creating a single management domain. The latter receives the high-level controls and commands generated from the End-to-End (E2E) management domain using domain integration fabric.

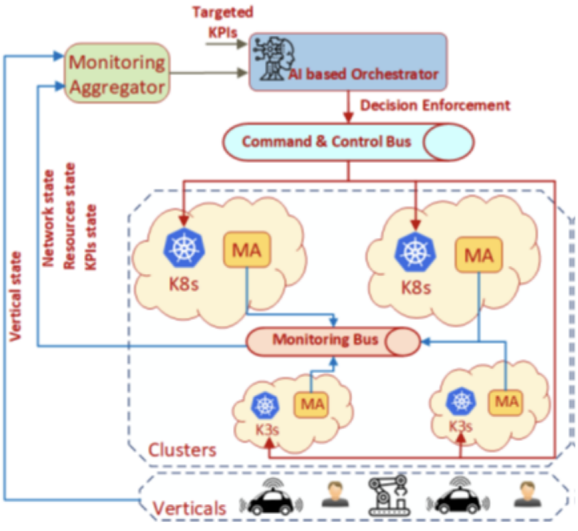

Figure 2: CHARITY's AIRO [11]

4.2 GPU-based primitives supporting AI-based service placement

Due to the continuous heavy charge that the previous item will take, a Machine Learning system to make the AIRO completely adaptative is needed, for which the project will leverage the use of GPU-based mechanisms.

4.3 Decentralized service replica management

Since we can consider that the system at-the-edge is made of entities with a specific geographic location representing a small-to-mid pool of potentially heterogeneous resource, but these resources can be repeated on that particular system-at-edge, a proactive and decentralized mechanism is needed to control the number of application replicas of the same service in an Edge Computing platform, while meeting the requested QoE promises.

4.4 Monitoring and prediction

Monitoring encompasses the process of collection, analysis and use of information systematically, that provides the continuous visualization and perception of the status and the progress of an application, service or infrastructure. Such continuous monitoring process provides a way to analyse the environment to check whether applications and infrastructure run as predicted. Indeed, the real-time monitoring of the environment allows for instance, to minimize the response time to incidents (e.g., the detection and mitigation of cyber attack). In past, this monitoring fundamentally served as a decision support for manual interventions of service and infrastructure administrators, but currently,as soon as something happens outside the expected behaviour, it is possible to take the appropriate actions and decisions. As we progress into more complex and challenging scenarios, monitoring and later, prediction algorithms, assume a whole new relevance in the orchestration and life-cycle management of next-generation applications. These mechanisms highly depend on a comprehensive real-time monitoring approach and on the quality of collected metrics.

4.5 Security and privacy-aware orchestration

As orchestration and scheduling of XR services form the core of the architecture, enclosing security for end-to-end service delivery becomes an important aspect. Apart from default security of the application and native cloud security, we achieve secured orchestration in three parts: (i) *Secured function execution:* we experiment secured function execution with TEEs (trusted execution environments) for improving privacy of the users accessing an application in a federated fashion. (ii) *Microservice security:* developed to find the minimum set of capabilities that containers need for executing their applications correctly while minimising their interactions with the OS kernel. This tool integrated with the orchestrator to perform both static and dynamic analysis of the microservices during task execution. (iii) *Learning with programmable switches:* Conventional switches in the network are getting replaced by programmable switches. We leverage these upcoming deployments to design and develop functions that can be deployed inside the switches to identify any anomalies during the orchestration.

5 Energy, data and computational-efficient mechanisms supporting dynamically adaptive and network aware services

5.1 New data services for AR and VR applications

The VR and AR UC applications of CHARITY demands ad hoc data services to achieve their goals. These services requires a significant R&D effort; the performance of existing solutions are insufficient to reach the UCs requirements and new algorithms have to be developed. In this section, we briefly describe such data services.

Holographic Assistant UC1.3 requires the efficient transmission of a large amount of 3D point cloud from the Cloud to the 3D Holographic device. Point cloud compression is an active research topic in Computer Graphics in this last years [12]. Despite this, the performance of the existing point cloud compression schemes, may be insufficient to guarantee the high amount of data required by this UC to provide an high-quality experience, and a custom encoder/decoder has to be developed. In the approach under investigation the point cloud is assembled into a voxel. Each voxel represents a point in the 3D space with additional information associated to it. Some of this information regards visibility information, the color, the transparency, and so on. This allows to control the resolution of the voxel according to the network and the computing resources available. Assuming the scene do not change quickly during the time, this volume of information can be treated as a 3D video and only the differences between one time step and the next one can be transmitted. The main idea is to encode such differences following approaches mutated by V-PCC [13].

The AR Collaborative Gaming UC3.1 needs a good synchronization between the real environment and the actions of the gamers. Two data services are used for this purpose, a *meshing service* (MS) which create a mesh starting from the 3D points extracted from the camera images, and another data service which estimates accurately the position and the orientation of the users, the *Localization Service* (LS). The MS enables the gamers to interact with the environment in an effective and robust way. In some cases this service is not required, for example for smartphones with 3D sensing capabilities, like the ones equipped with a Lidar. We built on image-based localization approaches to create the LS, in

particular on *structure-based* localization approaches, since the 3D reconstruction of the environment is available. In the specific context of this UC, the accuracy required is higher than the usual visual localization applications, but we do not suffer typical problems like weather or illumination changes, because we can assume that the game environment not change too much from the first acquisition. Modern end-to-end Deep Learning approaches [14, 15] provides good estimates, but do not generalize well. We will work on improving generalization exploiting the computation capability of the cloud to tune the networks on-the-fly and ensure high accurate position and orientation estimation.

5.2 Efficient storage and caching mechanisms

In edge computing, a large amount of data is generated and consumed by various edge applications. One of the key challenges in the development of applications at the edge is the efficient data sharing between the multiple edge clients. Data sharing can be realized within individual application frameworks or through an external storage service. Edge storage can greatly improve data access which in turn enables latency-sensitive applications. Despite the recent advancements in providing an edge storage solution, there are still issues left to be dealt with. Some issues related to the non-functional requirements of cloud-based application. In addition, edge nodes generally have limited computation, storage, network, or power resources and the distributed, dynamic and heterogeneous environment in the edge along with the diverse application's requirements poses several challenges.

To tackle these issues, one must leverage the core infrastructure and extend or integrate some of the most prominent software solutions such as MinIO, OpenStack Swift and CEPH with cloud-based storage services. However, it takes a more elaborate solution so as to deal with the inherent unreliability of the edge devices. Research for the efficient data placement takes a prominent role in developing a reliable edge storage solution with security coming in as a follow-up concern when heterogeneous storage systems in edge and cloud nodes need to exchange data. In addition, regarding resource management, several challenges concerning the adaptation to the dynamic environments and the large-scale optimization for the collaboration of multiple edge servers must be addressed. The literature presents multiple options regarding these topics. Near real-time decisions can be efficiently improved by moving the analytics "close" to the data. As a result, edge architectures can reduce the amount of data traversing the network, thus minimizing latency and overall costs. Among the most relevant work, there are a layered approach for data storage management and an adaptive algorithm which dynamically finds the trade-off between the quality and the amount of data stored at the edge and the cloud [16]. Regarding the question of which parts of the data to upload to the cloud, a distributed multi-level storage system model for edge computing is proposed in [17], which is based on a multiple-factors LFU (Least Frequency Used) replacement algorithm. In addition, caching at the edge can greatly improve data availability, retrieval robustness and delivery latency [18]. Therefore, efficient learning schemes for massive high-dimensional data are needed, in order to design efficient proactive caching algorithms. When it comes to security, the most popularly investigated option is the use of blockchains. Considerations regarding the use of blockchain technologies and tools for the implementation of an efficient edge storage system have been around for some time now [19]. The intention is to deal with issues such as reliability of the network and distribution of storage and computation over a large number of distributed edge nodes in a secure manner.

The guiding principle in CHARITY is to implement a hybrid distributed edge/cloud storage framework spread across heterogeneous edge and cloud nodes with intelligent decisions on data placement, data caching and considerations on performance (QoE) and security, emphasizing on the resolution of the problem of data distribution and offloading based on CHARITY's application requirements. As traditional LFU algorithms employed on the edge consider only the access frequency of the data, CHARITY will borrow ideas for the evaluation of data "importance" from various fields including fault tolerant distributed data stores, thus globally improving the storage's hit rate. Furthermore, CHARITY will provide optimized data prefetching and placement through intelligent admission mechanisms which are able to identify the correct time frame that data should be prefetched to the edge, preserving them as cache. Machine learning and predictive analytics mechanisms will be employed, aiming at a more concrete prediction model, optimizing the off-loading process by preventing bottlenecks and violations on QoS and QoE expectations of the platform. In addition, efficient machine learning-based schemes will be utilized in order to provide a new way of edge caching development. Finally, the integration of blockchain and edge computing would bring some benefits in terms of security, privacy and automatic resource usage.

5.3 Latency-sensitive and bandwidth-sensitive networking

In CHARITY, we aim at developing a dynamic multipath routing framework to improve the end-to-end communication in the context of the strict requirements of AR, VR, and holography-based applications. In order to so, it is essential to develop mechanisms which can facilitate the scheduling and routing of latency-sensitive and / or bandwidth-sensitive traffic. The component which shall be in charge of providing these functionalities is referred to as the Intelligent Traffic Routing mechanism. The Intelligent Traffic Routing mechanism leverages information regarding the various traffic

flows, the network topology and the network state, in order to establish traffic routing and scheduling functionalities in a manner which is compliant with the QoS requirements. The required information which relates to the traffic flows are their corresponding source, destination and QoS requirements. Furthermore, the Intelligent Traffic Routing mechanism shall also consume network traffic predictions which are provided by a dedicated Traffic Prediction mechanism. The Intelligent Traffic Routing Mechanism leverages the Software Defined Networking (SDN) paradigm in order to have access to vital information regarding the traffic and topology of the network. The SDN controller is able use Northbound APIs in order to establish communication with the application plane and Southbound APIs, such as OpenFlow, in order to communicate with the forwarding devices. These communication channels enable the SDN controller to examine the network state and flow-related information and then alter the flow tables of the forwarding devices accordingly. Furthermore, the Intelligent Traffic Routing Mechanism is designed to leverage Deep Reinforcement Learning (DRL) in order to conduct these functionalities in an optimal manner which is line with the QoS requirements. The centralized control provided by SDN greatly enhances the quality of DRL-based traffic engineering by enabling network policies to be centrally generated and then transferred to the forwarding devices. The formulation of the agent's Action Space is designed in a manner which is in accordance with the SDN paradigm.

Although there have been numerous scientific endeavours in regards to utilizing DRL-based paradigms in the context of SDN [20], only a few of them are designed to accommodate multipath routing while taking into consideration the QoS constraints [21]. CHARITY aims to expand upon the existing scientific literature [22] in regards to developing QoS-aware DRL-based structures which support multipath routing. To that end, the Action Space should be also modeled in a manner which can properly reflect the intricacies of multipath routing. On top of that, the State Space shall be implemented in a manner which includes the traffic predictions. By doing so, it is possible to enable the dynamic creation of policies that take into consideration the expected future state of the network as well as the ongoing one. Finally, the Intelligent Traffic Routing Mechanism shall also leverage Graph Neural Networks (GNNs) in order to enhance the efficiency of the DRL-based routing algorithms [23]. The use of GNNs shall enable the network structures to be represented in a more accurate way by properly encapsulating the intricate relations which are established among graph-based structures.

6 Application Management Framework

Along with the provisioning of advanced XR Service enablers, CHARITY has the goal to make such enhanced capabilities as accessible and usable by XR Application Developers as possible, to support improvements to the XR application development cycle, in terms of speed, cost and effectiveness. Application Management Framework (AMF) will be a component to enable the access to these capabilities; its baseline is the common CI/CD (Continuous Integration/Continuous Delivery) model, adapted and interpreted according to the needs and peculiarities of CHARITY, integrated with the development of custom tools. Network slice blueprint is a key concept of CHARITY and it's analogous to network service descriptor reported in [24]. AMF will enable XR Application Developers to design their own abstract network slice blueprints; these abstract descriptors will be converted in concrete descriptors by a component of the XR Service Orchestration layer, in order to create descriptors that are ready to be instantiated into runtime objects in CHARITY platform. Thus, AMF is the main entry point for XR application developers to define and handle their XR services at design time. Abstract slice blueprint handling can in general include the following:

- Design of abstract network slices, intended as blends of Virtual Network Functions into Network Services. The blend includes CHARITY provided artefacts (XR service enablers) and virtual links.
- Validation of generated network slice blueprints.
- Registration and storing in a common repository (XR Service Blueprint Templates Repository) of abstract network slice blueprints.
- Update of registered network slice blueprints.

Beyond the design time creation and management of abstract slice blueprints, AMF enables XR Application Developer to integrate their XR applications into the CHARITY platform via

- XR Application registration/onboarding. This part encompasses the uploading of application components in a repository shared with the Orchestration layer, accompanied by a proper abstract description;
- definition of application model templates describing the different application bricks, along with the description of their interconnection and interoperation;
- validation of composed applications in a segregated testing environment, according the tests written by XR Application Developers;

- management of dynamic changes to the application model during the application execution, including updates to running microservices, addition of new microservices, or decommissioning of running ones, keeping the consistency with the abstract application model repositories.

Each described artifact, that will need to be validated internally through unit and E2E tests, will be also validated with smoke tests, possibly along with native CHARITY components. Validation should be done in a test environment e.g.

- via single component smoke test run (if provided by XR Application Developers);
- via integration test provided by NextGen Application Developers, running with CHARITY components (mocks or full);
- security scan.

Both for abstract slice blueprint management and XR application definition and management capabilities, AMF components will use two approaches: loosely coupled interfaces with XR Service Orchestration layer to realize a publish/subscribe pattern and ome shared repositories to store artifacts created by AMF and vice versa to read XR enablers descriptors.

7 Plans for showcasing and validation

The showcasing and validation activities of CHARITY intend to demonstrate the feasibility of the CHARITY framework to support the deployment and orchestration of next generation XR services using a more autonomous and intelligent approach. Nevertheless, the entire CHARITY framework is composed by a multitude of different components and mechanisms, which, from an integration standpoint, poses numerous challenges.

First, CHARITY addresses the research and development activities of several innovative components which are not available (or mature) but are critical to fulfil the requirements of the advanced media scenarios envisioned. The different maturity levels of concepts and mechanisms proposed by CHARITY pose a significant challenge on how to validate the CHARITY framework as a whole.

Then, CHARITY considers a new class of next generation of XR services designed to maximize the benefits of distributed, as Service Oriented Architectures (SBA) and zero-touch approaches. For instance, this includes cloud-edge continuum scenarios spanning across distinct domains (both administrative and technological) with all the inherent computing and network requirements of immersive (and real-time) applications. Not only showcasing but performing a comprehensive evaluation and validation of these complex and heterogeneous scenarios is in itself a challenging activity.

In that sense, as part of the validation and showcasing efforts, we started to outline the technical integration plans. These plans comprise the design of an overall integration and evaluation framework, the identification of clusters of key components and cross layer aspects which are of utmost relevance for understanding the feasibility of the integrated platform (e.g. which and how the different elements spreading over different domains can be fully integrated and assessed) and common collaborative tools which shall be used to organize the geographically distributed development teams.

Then, as part of the integration and validation plans, we started by identifying existing open sources and enablers, and map them with the components of CHARITY framework. Namely, built upon the idea that CHARITY should support such multi-domain scenarios, we initially investigated the usage of cloud computing and cross domain tools as a more unified way to expose APIs and resources at several layers (i.e., VMs, containers, VNFs) on the envisioned Cloud Native environment. Such an earlier survey of existing tools, initially focused on deployment and provisioning of XR services, allow us to better plan the overall integration environment and incrementally understand how to integrate the various CHARITY elements. Moreover, built upon the idea that CHARITY should support such multi-domain scenarios, such integration strategy of supporting distinct domains, has the benefits of allowing us to leverage existing partners infrastructures as part of the whole integration environment. For instance, a hybrid set of different cloud computing platforms (e.g., OpenStack deployments on partners premises, public cloud domains) can be used to form the whole (multi-domain) integration environment where different partners can deploy, validate and showcase their proposed mechanisms and systems.

8 Conclusions and Outlook

We have presented the overall approach and vision of the CHARITY project toward supporting next-generation XR applications, exploiting a combination of networking, storage, compression and dynamic management techniques in

order to provide XR service deployment and adaptive execution over heterogeneous Cloud and Edge networking and computing resources.

The future importance of XR application in their different characterization of 3D VR, AR, and holographic interfaces cannot be overstated. The project use-cases we presented in Sect. 2 are a key sample, even if a small one, of the multitude of applications patterns that the CHARITY architecture and technology will enable. Future network-mediated XR interaction will enhance the effectiveness and reliability of remote collaboration, improving all human processes that exploit distanced interaction, as well as enable new ones.

The CHARITY approach supports applications via the AMF abstraction of network slice blueprints, to be dynamically enacted by the platform. The modularity and separation of concerns that the approach provides allow exploiting the sophisticated techniques described in Sections 4 and 5 and reaping the advantages they bring in terms of efficiency, QoS and QoE, while minimizing the effort on part of the application developer and improving the cost and reliability of XR.

CHARITY, despite being in its initial stage, has already advanced quickly in defining its overall platform design. We are investigating and developing technologies, algorithms and SW architectures to support the initial design, and we expect several advances with respect to the state of the art to emerge from project activities. This holds specifically for the research efforts about strategies to maximize the exploitation of available bandwidth, as well as those about dynamic, adaptive techniques to allocate and manage network and computation resources.

There is clearly a positive interaction among the research aspect and the practical issues of industrial application developers. This cooperation and synergy, besides backing up the development of a friendlier XR support API and its experimental validation, is pushing for new advances in several of the research fields that CHARITY activities entail and were thus outlined in this paper.

Acknowledgment

This work is part of the CHARITY project that has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 101016509.

References

[1] Wikipedia contributors, "Pepper's ghost — Wikipedia, the free encyclopedia," accessed 15-Oct-2021. [Online]. Available: https://en.wikipedia.org/wiki/Pepper%27s_ghost

[2] R. Häussler, Y. Gritsai, E. Zschau, R. Missbach, H. Sahm, M. Stock, and H. Stolle, "Large real-time holographic 3D displays: enabling components and results," *Appl. Opt.*, vol. 56, no. 13, pp. F45–F52, May 2017. [Online]. Available: http://www.osapublishing.org/ao/abstract.cfm?URI=ao-56-13-F45

[3] S. A. Benton and J. Bove, V. Michael, *Holographic Imaging*. USA: Wiley-Interscience, 2008.

[4] E. Zschau, R. Missbach, A. Schwerdtner, and H. Stolle, "Generation, encoding, and presentation of content on holographic displays in real time," in *Three-Dimensional Imaging, Visualization, and Display 2010 and Display Technologies and Applications for Defense, Security, and Avionics IV*, B. Javidi, J.-Y. Son, J. T. Thomas, and D. D. Desjardins, Eds., vol. 7690, International Society for Optics and Photonics. SPIE, 2010, pp. 118 – 130. [Online]. Available: https://doi.org/10.1117/12.851015

[5] M. Hassandra, E. Galanis, A. Hatzigeorgiadis, M. Goudas, C. Mouzakidis, E. M. Karathanasi, N. Petridou, M. Tsolaki, P. Zikas, G. Evangelou *et al.*, "A virtual reality app for physical and cognitive training of older people with mild cognitive impairment: mixed methods feasibility study," *JMIR serious games*, vol. 9, no. 1, p. e24170, 2021.

[6] G. Papagiannakis, P. Zikas, N. Lydatakis, S. Kateros, M. Kentros, E. Geronikolakis, M. Kamarianakis, I. Kartsonaki, and G. Evangelou, "Mages 3.0: Tying the knot of medical vr," in *ACM SIGGRAPH 2020 Immersive Pavilion*, 2020, pp. 1–2.

[7] P. Zikas, M. Kamarianakis, I. Kartsonaki, N. Lydatakis, S. Kateros, M. Kentros, E. Geronikolakis, G. Evangelou, A. Apostolou, P. A. A. Catilo *et al.*, "Covid-19-vr strikes back: innovative medical vr training," in *ACM SIGGRAPH 2021 Immersive Pavilion*, 2021, pp. 1–2.

[8] M. Kamarianakis, N. Lydatakis, and G. Papagiannakis, "Never 'drop the ball' in the operating room: An efficient hand-based vr hmd controller interpolation algorithm, for collaborative, networked virtual environments," in *Advances in Computer Graphics*, N. Magnenat-Thalmann, V. Interrante, D. Thalmann, G. Papagiannakis, B. Sheng, J. Kim, and M. Gavrilova, Eds. Cham: Springer International Publishing, 2021, pp. 694–704.

[9] ETSI GS ZSM 002, "Zero-touch network and Service Management (ZSM): Reference Architecture," European Telecommunications Standards Institute (ETSI), Tech. Rep., Aug. 2019.

[10] T. Taleb and e. al., "Towards supporting xr services: Architecture and enablers," *Submitted to IEEE IOT Journal*.

[11] A. Boudi, M. Bagaa, P. Pöyhönen, T. Taleb, and H. Flinck, "Ai-based resource management in beyond 5g cloud native environment," *IEEE Network*, vol. 35, no. 2, pp. 128–135, 2021.

[12] C. Cao, M. Preda, and T. Zaharia, "3d point cloud compression: A survey," in *The 24th International Conference on 3D Web Technology*, ser. Web3D '19. New York, NY, USA: Association for Computing Machinery, 2019, p. 1–9.

[13] D. Graziosi, O. Nakagami, S. Kuma, A. Zaghetto, T. Suzuki, and A. Tabatabai, "An overview of ongoing point cloud compression standardization activities: video-based (v-pcc) and geometry-based (g-pcc)," *APSIPA Transactions on Signal and Information Processing*, vol. 9, 04 2020.

[14] A. Kendall, M. Grimes, and R. Cipolla, "Posenet: A convolutional network for real-time 6-dof camera relocalization," in *2015 IEEE International Conference on Computer Vision (ICCV)*. Los Alamitos, CA, USA: IEEE Computer Society, dec 2015, pp. 2938–2946.

[15] A. Valada, N. Radwan, and W. Burgard, "Deep auxiliary learning for visual localization and odometry," in *2018 IEEE International Conference on Robotics and Automation (ICRA)*, 2018, pp. 6939–6946.

[16] I. Lujic, V. De Maio, and I. Brandic, "Efficient edge storage management based on near real-time forecasts," in *2017 IEEE 1st International Conference on Fog and Edge Computing (ICFEC)*. IEEE, 2017, pp. 21–30.

[17] J. Xing, H. Dai, and Z. Yu, "A distributed multi-level model with dynamic replacement for the storage of smart edge computing," *Journal of Systems Architecture*, vol. 83, pp. 1–11, 2018.

[18] Y. Huang, X. Song, F. Ye, Y. Yang, and X. Li, "Fair and efficient caching algorithms and strategies for peer data sharing in pervasive edge computing environments," *IEEE Transactions on Mobile Computing*, vol. 19, no. 4, pp. 852–864, 2019.

[19] R. Yang, F. R. Yu, P. Si, Z. Yang, and Y. Zhang, "Integrated blockchain and edge computing systems: A survey, some research issues and challenges," *IEEE Communications Surveys & Tutorials*, vol. 21, no. 2, pp. 1508–1532, 2019.

[20] C.-C. Fang, C. Cheng, Z. Tang, and C. Li, "Research on routing algorithm based on reinforcement learning in sdn," *Journal of Physics: Conference Series*, 2019.

[21] M. K. Awad, M. H. H. Ahmed, A. F. Almutairi, and I. Ahmad, "Machine learning-based multipath routing for software defined networks," *Journal of Network and Systems Management*, vol. 29, pp. 1–30, 2021.

[22] J. Rischke, P. Sossalla, H. Salah, F. H. P. Fitzek, and M. Reisslein, "Qr-sdn: Towards reinforcement learning states, actions, and rewards for direct flow routing in software-defined networks," *IEEE Access*, vol. 8, pp. 174 773–174 791, 2020.

[23] P. Almasan, J. Suárez-Varela, A. Badia-Sampera, K. Rusek, P. Barlet-Ros, and A. Cabellos-Aparicio, "Deep reinforcement learning meets graph neural networks: exploring a routing optimization use case," 2020.

[24] ETSI, "Network functions virtualisation (nfv) release 4; management and orchestration; network service templates specification (etsi gs nfv-ifa 014 v4.2.1)," 2021, last accessed 11 October 2021. [Online]. Available: https://www.etsi.org/deliver/etsi_gs/NFV-IFA/001_099/014/04.02.01_60/gs_NFV-IFA014v040201p.pdf