Towards 6G-based Metaverse: Supporting Highly-Dynamic Deterministic Multi-User Extended Reality Services

Hao Yu, Masoud Shokrnezhad, Tarik Taleb, Richard Li, and JaeSeung Song
Oulu University, Finland; Email: {firstname.lastname}@oulu.fi
Futurewei Technologies, USA; Email: richard.li@futurewei.com
Sejong University, Korea; Email: jssong@sejong.ac.kr

***Abstract*—Metaverse is the concept of a fully immersive and universal virtual space for multiuser interaction, collaboration, and socializing; forming the next evolution of the Internet. Metaverse depends on the convergence of multiple broad technologies that enable eXtended Reality (XR), which is an umbrella term for technologies that lie on the reality–virtuality continuum, namely virtual reality (VR), Augmented Reality (AR), and mixed reality (MR). Compared to streaming volumetric content to a single user, the XR applications that involve multiple users who simultaneously watch the same volumetric content (e.g., VR-based online training/education and multi-user online gaming) are much more attractive. Multi-user XR/Metaverse puts additional demand on the underlying networks, making it more difficult to offer high-quality immersive material in real-time. To cope with this, this article presents a comprehensive system and component design for immersive and seamless multi-user XR experiences. To satisfy the Quality of Experience/Quality of Service (QoE/QoS) requirements and especially stream synchronization requirement in XR collaboration scenarios, we propose an AI-powered deterministic multi-user extended reality resource orchestrator (PRECISENESS) that aims to solve the multi-user XR service provisioning problem. Finally, we demonstrate the performance of our proposed solution in a single-site multi-user XR use case. The obtained results demonstrate that our solution can deliver high-quality immersive XR services.**

***Index Terms*—Multi-user XR, Semantic Networking, Deterministic Networking, Holographic-type Communication, Haptic Communication, Immersive Services, Metaverse, Beyond 5G, and 6G.**

I. Introduction

The Metaverse, as a new application of a blended shared space whereby users can interact with each other in physical/virtual environments and extended reality (XR) in a seamless, immersive and interactive way, has drawn huge attention in both industry and academia. As shown in Fig. 1, by leveraging XR technologies, Metaverse can change the way people work, learn, entertain and socialize, for instance, in sightseeing classroom education, and interactive gaming areas. The term ”extended reality” (XR) refers to a broad category of reality that makes use of technologies that provide an immersive experience in order to construct digital worlds in which data are represented and projected. XR encompasses a wide range of subfields, most notably virtual reality (VR), augmented reality (AR), and mixed reality (MR).

XR platforms commit to provide immersive user experience by creating distinct artificial digital environments in which users feel immersed and behave in the same manner as they would in real life or superimposing digital things into the real world to provide consumers access to more interesting information. To support immersive user experience, volumetric videos, formed with six degrees of freedom (6DoF), which includes the position (X, Y, Z) and the orientation $(yaw, pitch, roll)$ of the viewer, have contributed to the rise in popularity of this type of videos in recent years. When watching a volumetric video, users have the ability to freely select any viewing angle of the 3D scene that they prefer, giving them an experience that goes beyond 360-degree video constrained to 3 degrees of freedom. In addition, a multimodal platform takes advantage of the multiple human sensory channels in a virtual domain, allowing for diverse types of interactions, so called Tactile Internet (TI). This will also contribute to a better user experience

In contrast to conventional video-on-demand (VoD) streaming, XR applications usually involve the live volumetric streaming, that is, the video is transmitted in real time directly from the device that originated the stream to the device that will play it. This occurs without the video first being stored on a server. Live video streaming is more vulnerable to delays in the network than VoD streaming. For conventional live streaming videos, the live system needs to handle a huge volume of videos from broadcasters to thousands of viewers located all over the world in a very short end-to-end latency (e.g., 100 ms [1]). In addition, panoramic live video streaming enables a flexible interaction, meaning that the user can walk around and expect to see different perspectives of the panorama. As a result, the latency (e.g., motion-to-photon latency) and its variance have been made even more stringent so that the display may be updated without causing any motion sickness. This latency should not be more than a few milliseconds in order to ensure a seamless user experience [2].

As for VR/AR live streaming, we should synchronize the video streams from/to multiple users not only from a multimedia perspective, but also from a game engine perspective. On the one hand, XR applications, requiring devices with screens and computing components, have the same requirements as those for conventional live videos and panoramic live videos. On the other hand, they also need sensors (e.g., motion sensors), equipped on the XR device, tracking the users'

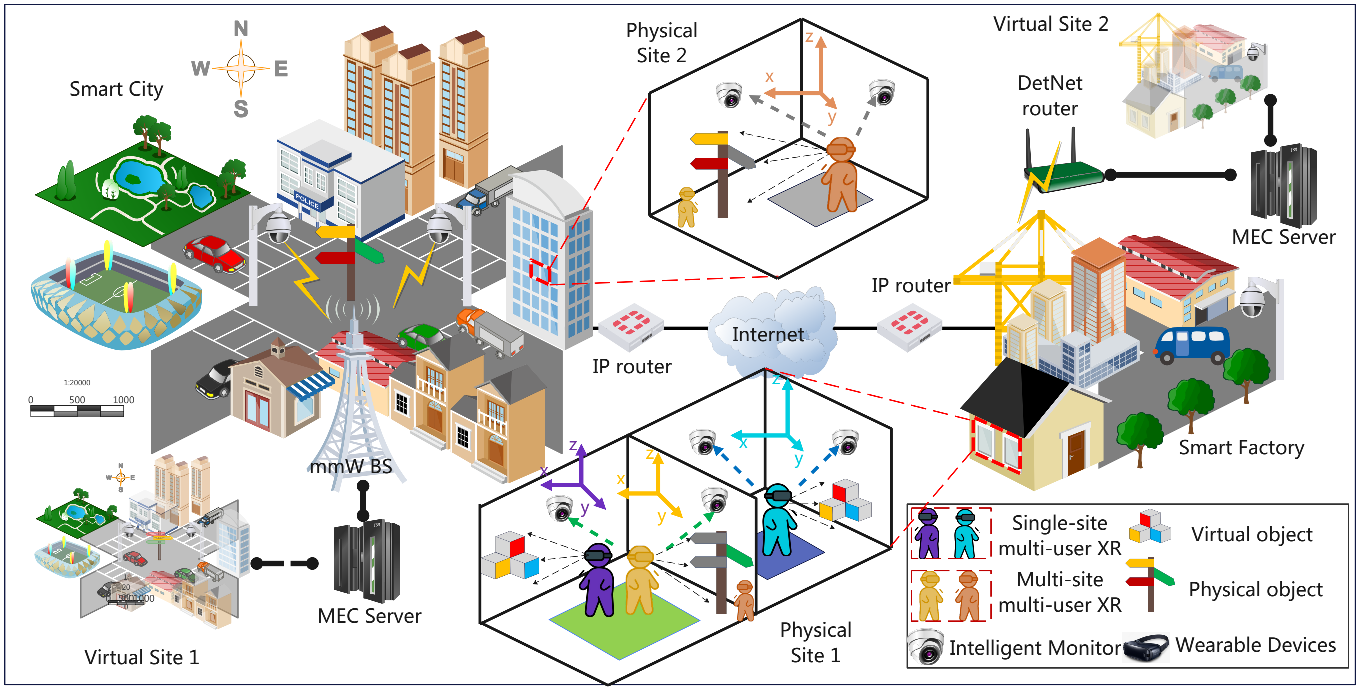

Fig. 1. Different communication scenarios for Multi-user XR enabled Metaverse.

behaviours to support the interaction between the multiple users in the virtual/physical space. Although modern computer game engines have many mechanisms to optimize delays and reduce lags to allow object state synchronization (e.g., multiple players in Multiplayer Online Battle Arena (MOBA) games, such as League of Legends), they typically only deal with very limited user interaction, and thus may fall through when it comes to real-time capturing and transmission of data with multiple modalities. For example, the transmission of audible words and the movement of the lips should be synchronized to render a proper presentation in an XR teleconference. Also, the streams should also be synchronized if multiple performers are located in different places to present an XR concert together. Note that, despite the similar designs, VR and AR accomplish very distinct tasks in very different ways. Augmented reality augments your vision with a real-world setting while virtual reality replaces it completely, more specifically, AR users can control their presence in the real world; VR users are controlled by the system. For instance, AR has typically less realistic content requirements than VR, which needs vivid information to fully immerse people in the virtual environments. But both AR and VR applications need to accommodate a certain amount of user interaction. In contrast to displaying images or playing movies, user interaction, for example, waving a hand right or left to switch content, pressing a finger forward to push a “button” and saying “close” to stop the application, requires real-time communications between user/user or user/server. Therefore, the deterministic networking (DetNet) [3] capabilities, which enable deterministic data delivery, processing and synchronization capacity provided by the underlying networks, become necessary to support the high-quality immersive user experience in real-time.

In this paper, we propose an artificial intelligent (AI)-emPoweRed dEterministiC multI-uSer ExteNded rEality reSource orcheStrator (PRECISENESS) for the deterministic service provision of multi-user XR applications. With the synergy of AI and deterministic networking (DetNet) technologies, the proposed PRECISENESS can solve the XR service provision problem with constraints of different types of stream synchronization in an efficient and optimal way. On one hand, AI-based decision engine and analyze engine, in alignment with the concept of adaptive AI, will be used to capture, model and analyze the user and environment dynamics over time and continuously re-trained with new data and conditions. On the other hand, a monitoring system, that will collect the service and network metrics, will be leveraged to facilitate the deterministic multimedia stream operation and management.

The remainder of this article are organized as follows. We first briefly review the emerging enabling technologies which can facilitate the advancement of 6G, XR and Metaverse. Secondly, we provide an overview of system and component design supporting immersive and seamless multi-user XR experience. Next, we will analyze the key XR Quality of Experience (QoE) requirements and their Key Performance Indicators (KPIs) on the underlying networks. By introducing two potential use cases in multi-user XR, we shed light on the concept of inter/intra-media synchronization, inter-source/destination synchronization for XR collaboration scenarios. As a further step, we propose PRECISENESS to solve the multi-user XR service provisioning problem. Finally, we perform a case study to show the performance of our proposed solution in a single-site multi-user XR case. The obtained results show that our solution has great potential in delivering high-quality immersive XR services.

II. Enabling Technologies for multi-user XR

Diverse research works of recent literature have concentrated on the goal of improving the performance of communication networks. Conventional technologies have been enhanced, and novel methods and technologies have been developed. The following technologies are among the most prominent ones in relation to the Metaverse and XR-centered applications:

A. 5G and Beyond

5G has promised the provisioning of high capacities at the radio access network (RAN) level. Beyond-5G is expected to provide higher RAN capacities. With such high-capacity access technologies, users will be able to connect through the Metaverse in a more natural and intuitive manner, with less lag and a higher level of interactivity. These technologies will provide connectivity to a greater number of connected User Equipment (UEs) and will set the stage for supporting a large number of anticipated Metaverse users. The performed connection functions will be extremely secure and reliable, which will be vital for protecting the privacy and integrity of user data in XR applications. Moreover, beyond-5G networks will include more advanced capabilities, such as dynamic, real-time, on-demand radio access network slicing, which will enable the formation of virtual networks that can be configured to fit the requirements of various types of applications in the Metaverse, notably to isolate and safeguard XR streams. This will enable Metaverse developers to construct virtual sub-networks that are more specialized, efficient, and high-performing to enable their immersive and responsive applications.

B. Compute-First Networking (CFN)

Through real-time monitoring of computing and networking resources, the CFN system is able to immediately adapt to any changes in accessibility, availability, or quality of computing capacity in the networks caused by dynamic traffic patterns or restrictions imposed by administrative authorities in each domain or region [4]. In addition, collaborative orchestration of the resources of other domains will allow the system to allocate the resources of one domain based on the real-time status of the others. Therefore, allocations of a portion of resources that result in an overload in another portion of them will be avoided, and the system will be tailored in a manner that it balances the load of both cloud/edge and network resources, reduces the overall cost, or minimizes energy consumption. As a result, more resources can be activated and their utilization will increase, serving a significant number of Metaverse and XR users, who can be wirelessly connected via emerging beyond 5G access technologies [5].

C. Semantic Networking

In the majority of traditional packet-based systems, each packet is considered an irreducible, independent, and self-sufficient unit that is identified by a pre-defined specific destination address and is treated according to underlying congestion conditions and global or local configurations. This paradigm is advanced by semantic networking [6], which focuses on the WHAT of requests rather than the WHERE (the final destination). In this concept, network devices are configured with rules that are semantically aware, and these rules establish a route for each semantic toward its final destination. A number of different factors, such as service class, geographic location, cloud hardware specification, QoS class, and payload type, could be used to define the semantics. Now, instead of controlling requests individually, the focus will be on a set of semantics, and consequently, a relatively smaller set of rules will need to be adjusted based on resource changes or end-user dynamics.

D. Adaptive AI

Recent developments in Deep Reinforcement Learning (DRL) have exhibited amazing efficiency in a range of tasks, although they typically rely on an agent that specializes in learning a specific job of interest. In addition, following every substantial change, RL agents usually require further training to adapt to the new situation, and even after receiving this training, they lack the ability to generalize to new variants, even for simple problems. Therefore, dynamic environments need the development of unique learning processes, the once-in-a-lifetime train model mindset should be replaced by Adaptive AI, a novel approach wherein models are continuously retrained with new data and conditions. Continual Learning (CL) is concerned with the adaptation of the RL agent to the evolution of these environments over time [7]. CL-based approaches are intended to address two major challenges: 1) catastrophic forgetting, which refers to the loss of performance on old tasks after learning new tasks, and 2) interference, which occurs when two tasks have incompatible (or even contradictory) optimal actions for the same observation.

It is anticipated that Adaptive AI will be one of the most important enablers for the delivery of emerging services such as Metaverse and XR applications [8]. In these applications, a set of UEs with ever-changing QoS/QoE requirements compete for network and computing resources. Due to the fact that these UEs are mobile and can be constantly moved from one geographical location to another at high speeds and frequencies, the number and type of UEs with data to transmit over the access channels may fluctuate over time. In addition, requests and traffic patterns may vary from moment to moment, either within a single UE or across multiple UEs in terms of the active services-seeking connection. Likewise, the conditions of resources are susceptible to change due to a vast array of external factors. In order to realize future self-sustaining Metaverse use-cases, especially for XR-related scenarios, adaptive decision-making algorithms are essential.

III. Multi-user XR system overview

As shown in Fig. 1, Metaverse is illustrated as a utopian convergence of various virtual environments which are connected to facilitate social events. It naturally involves multiple users located in different parts of the world. Therefore, corresponding data and stream synchronization technologies are

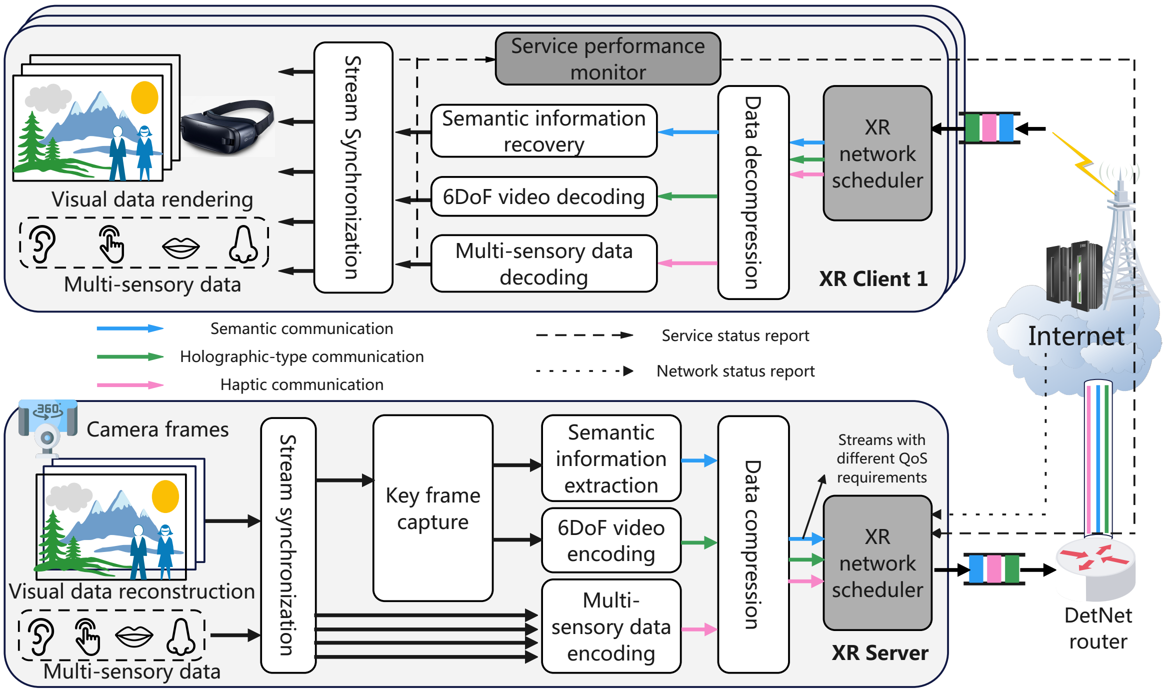

Fig. 2. Illustration of components for the multi-user XR system.

needed to support users' multi-user XR collaborations. In this section, we introduce the multi-user XR collaboration concept and then illustrate the essential components in the multi-user XR system.

A. Multi-user XR collaboration

Multi-user XR collaboration refers to bringing together groups of people for remote activities, such as meetings, conferences, design reviews, and classroom sessions through the use of XR devices and technologies. Individuals and organizations can now communicate in a much more visceral and connected manner, engaging a larger sense of physical presence. For single-user XR stream, when several correlated media components are involved in a multimedia system, e.g., video/audio stream or multi-sensory data stream, different media frames captured at the same time must be simultaneously presented at the corresponding output devices, that is, different media streams must be aligned in time dimension. For instance, during the speech of a remote user, it is important to synchronize the audible words with the related movement of the lips in order to properly communicate. This is typically known as *lip synchronization*. Furthermore, interacting with other people requires greater efforts, e.g., XR users from different sites may operate on the same virtual object in one XR scene. Specifically, media streams, originated from different sources or delivered by different senders, must be synchronized. A typical example is the synchronization between video streams sent by different video servers (e.g., in surveillance systems). In addition, the synchronization between different media components, from either the same or different sources, which are delivered through different paths or via different (e.g., wireless or wired) technologies or networks, should be also investigated. More details about the inter/intra-media, intra/inter-user stream synchronization will be discussed in Section V.

B. Detailed architecture for multi-user XR system

To support the seamless, immersive multi-user XR services, XR system architecture and components must be designed carefully by coordinating the XR services and underlying networks to meet their strict QoS/QoE requirements.

Our proposed end-to-end architecture is illustrated in Fig. 2. At the XR server, after being captured, different media streams will be synchronized in advance for further processing. Key video frames will be captured for a) key semantic information extraction for reducing the unnecessary transmission of duplicate contents, e.g., static background; b) object viewing angle selection according to the point of view that client user presents, to enable the semantic networking paradigm. Then, multiple types of media data will be compressed and then streamed to the underlying networks. Before being emitted into networks, the streams (i.e., semantic stream, video stream and sensory data stream) will be prioritized by an intelligent network scheduler according to the QoS requirement levels of different types of streams. For example, the semantic information and part of video data change slowly over time compared with the tactile data (e.g., motion information) or audio data, which can be labeled as low priority level. By incorporating service performance monitors at the XR client, the goal of this system is to provide XR services with error resistance and to achieve the highest possible level of QoS. The purpose of client-side network performance monitors is to monitor the network parameters, such as bandwidth, loss rate, as well as network round-trip times. These parameters are communicated in a periodic fashion to the XR network scheduler which is located at the server. At runtime, once the feedback from the client is passed to the intelligent network scheduler, it will then adjust the service parameters related to priority level, source data rate or routing path for the stream(s) that suffer from the high end-to-end latency, jitter or packet loss, etc. In addition, Adaptive AI will be applied in the scheduler to capture the ever-changing QoS/QoE requirements and dynamic XR users, ultimately realizing the self-sustaining XR services. At the same time, a close-loop network management solution should be provided under compute-first network (CFN) framework to discover the available compute capability, optimize the utilization of compute resources for XR stream processing, including compressing/decompressing, encoding/decoding, semantic information extraction, etc., and perform the lifecycle management for XR services.

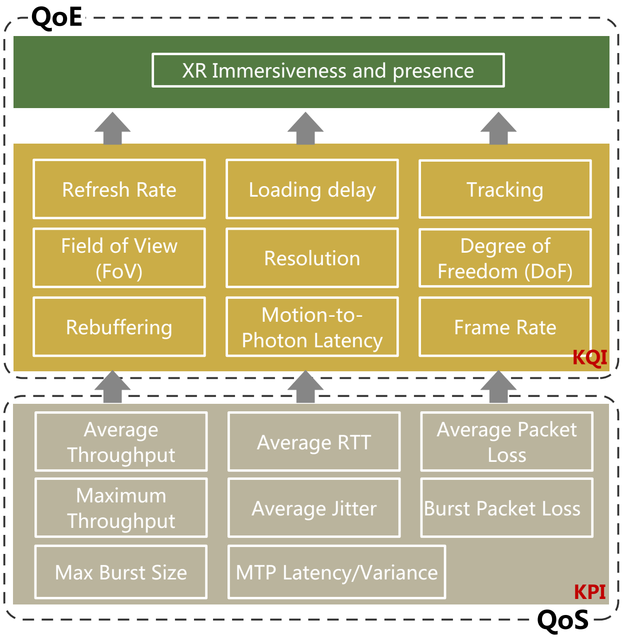

Fig. 3. KPIs to determine the QoE of XR users.

IV. Quality of Experience (QoE) and Quality of Service (QoS) requirements for XR communications

Being delivered via the Internet from the XR cloud (i.e., host) to the XR headsets (i.e., clients), the VR streaming has different requirements, including those related to the application parameters (e.g., video quality, frame rate, and resolution) [9] [10], which also closely relate to the physical characteristics of human perception and network parameters (e.g., throughput, RTT latency, packet loss rate, and reliability). Although Quality-of-service (QoS) models have been proposed to capture the qualitatively or quantitatively defined performance agreement between the service provider and the user applications when evaluating aspects of system performance, such as network conditions and video quality, they are incapable of measuring the user's satisfaction with interactive multimedia applications and devices. Quality of Experience (QoE) models have been extended from the network-centric QoS for multimedia systems to explain the actual quality of the formation process and reflect "the degree of delight or annoyance of the user of an application or service" by developing subjective and objective quality metrics. The section will introduce some relevant Key Quality Indexes (KQIs) that will effect QoE and the corresponding required QoS KPIs.

A. *Quality of Experience*

The quality of experience is a measure to quantify the overall users' subjective satisfaction with a service, as shown in Fig. 3. There are several key quality indices (KQI) defined in 3GPP TR 26.928, which can be used to evaluate the users' satisfaction on XR experience from different perspectives, e.g., media quality index, interaction quality index, and performance quality index, to infer the XR QoE [11].

Frame rate: The frame rate represents the rate at which frame-based images are displayed continuously on a display. The frame rate of XR content must be compatible with the display device's frame rate attribute. In addition, VR game applications have higher frame rate requirements because images are rendered by Graphics Processing Units (GPUs) after being captured by cameras. Low frame rate or video pauses are one of the causes of vertigo when using VR services. AR usually has lower requirements regarding to the level of content realism than VR [12].

Resolution: A video with a higher spatial and temporal resolution seems smoother and brighter than one with a lower quality. For instance, 3D models provision with higher polygon counts or 8K resolution (8196 x 4096) per eye achieved can eliminate the pixelation completely. A sustained frame rate of 90Hz and beyond is required to truly generate sense of immersiveness and presence. Low resolution and low pixels per inch (PPI) can cause pixelation and make the user feel as if they are viewing through a screen door. A greater bitrate is typically accompanied with a higher resolution.

Field of View: The FoV measures the range of visual environments at any given time. The level of FoV, which represents the extent of the observable environment, is one factor that contributes to the uniqueness of 360 video experiences. A broader FOV could contribute to an enhanced sense of immersiveness and presence. Consequently, the FoV of the HMD is an important metric for determining the extent to which a VR device could contribute to the creation of an immersive experience.

MTP Latency: Motion-to-Photon (MTP) is defined as the delay between a user action and its consequences on the display. In mobile networks, ensuring a low MTP latency with a high visual quality is the primary challenge. High MTP latency values transmit inconsistent signals to the vestibulo ocular reflex (VOR), which can result in motion sickness. This delay should be less than 20 milliseconds for smooth virtual space movement [13]. Reduction in time delay is an essential aspect of QoE for live streaming, but is considerably less crucial for other types, such as Streaming Video on Demand.

Tiling Artifacts/Mosaic: For evaluating the fluidity of cloud VR games, tiling artifacts/mosaic is a critical factor because users perceive mosaics in certain areas of the game image. In VR gaming, if some video frame information (some block information in the video frame) is lost, the image can still be displayed, but mosaics appear in some areas, disrupting the fluidity of the VR gaming experience for the users.

Multimodal Experience: The XR system should enable people to perceive and control actual and virtual objects in Metaverse, e.g., real-time physical tactile experiences, thermoception and olfactory perception, by incorporating Internet of Things (IoT) devices into AR/VR systems [14]. In smart hospitals, for instance, telesurgery should allow doctors to feel the sense of the touch and force feedback via haptic clothing/equipment. Consequently, the communication channel must ensure real-time haptic data delivery (i.e., touch, actuation, motion, vibration, and surface texture) [15].

B. *Quality of Service*

Compared with the KQIs, which are designed to quantify the feeling from human's perspective, the KPIs on QoS evaluate the network performance from the engineering system perspective.

High Bandwidth Requirement: The main parameter to determine the XR content quality includes pixels per degree (PPD), color depth, and frame refresh rate. PPD is related to the display's resolution indicating the pixel number per degree. Color depth indicates how many bits to distinguish a color's grayscale, whilst the frame refresh rate is defined as the number of refreshing frames per second. For a realistic view, 120 frames per second are required. Assuming resolution is 2K (i.e., 2160*1200), a colored pixel is represented by 8 bytes and taking into account a maximum video compression ratio of 1:600 (using H.265 HEVC codec), the VR system would require a bitrate of up to 3*8*120*2160*1200/600=12,441,600 bits/s ≈ 12 Mbps to ensure such video quality.

Low Latency Requirement: The end-to-end latency can be modeled as the sum of sensing time, rendering time, streaming time, and displaying time [10]. Sensing time is typically 400 microseconds for the headset and sensors to gather measurements, mostly for localization and controller inputs. The rendering time for both the foreground and background is between 5 and 11 milliseconds, depending on the complexity of the virtual environment [11]. To provide ultra-low latency, which is imperative for real immersive experiences, many transport layer protocols promote reliable and fast data delivery. For example, Real-Time Media Protocol (RTMP) is the most widely adopted protocol among TCP-based solutions. Despite efforts to strengthen the transport layer for real-time communications, the latency controls needed for extremely precise granularity at the packet or frame level should be also investigated. Cross-layer design should be performed to support extreme deterministic communications from underlying network infrastructure.

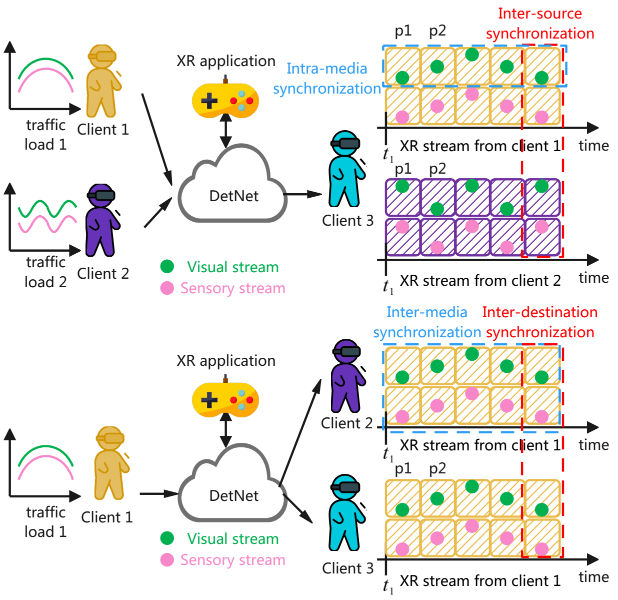

Fig. 4. Different stream synchronization types for multi-user XR.

V. Stream Synchronization for Multi-User XR

Besides the bandwidth and latency requirements of XR services, we should also pay more attention to the stream synchronization requirements in the XR collaboration scenarios. Fig. 4 illustrates the concept of multiple types of streams applied to a unidirectional XR service. In the illustration, video and multi-sensory streams from an XR game application are being played by a collection of dispersed receivers. In this case, no matter if it is media stream or user interaction data, *intra-media synchronization* must be provided for each of the involved streams to provide proper and natural playout processes, such as the evolution of a video sequence displaying a virtual object in motion. Specifically, if the media source takes a video sequence at 40 frames per second, each frame must be delivered and shown at the receiver side within 25 milliseconds. When multiple correlated media components are included in an XR system, the temporal dependencies between their streams must be also maintained during playout. This is known as *inter-media synchronization*. However, this is difficult to accomplish when the various media components have disparate features (e.g., media type, and traffic load) and requirements (e.g., processing, bandwidth, and latency). Similarly, the *intra-media synchronization* process might have an effect on the *inter-media synchronization* processes for specific media components. Therefore, a good trade-off and coordination must be provided to satisfy both requirements and ensure adequate QoE levels.

Inter-media synchronization may also involve the synchronization of media components originating from distinct sources or sent by distinct senders. This particular subtype of *inter-media synchronization* is commonly known as *inter-source synchronization* (see Fig. 4(a)). The synchronization of video streams supplied by separate video servers is a common example. In addition, it is possible for distinct media components from the same or separate sources/senders to be provided via different routing paths, or even via different (wireless and/or wired) technologies or network domains with the same or different characteristics.

Single or many receivers/devices are capable of synchronizing the playback of the associated media components, regardless of their format, number, or mix. The associated receivers/devices playing the same or related content may be geographically close together (e.g., in the same local networks) or geographically dispersed (e.g., in different buildings, cities, or countries). Multiscreen settings and structures with multiple distributed loudspeakers are examples of the former circumstance. In such circumstances, the absence of *inter-receiver synchronization* will result in disjointed multiscreen experiences and the sense of irritating echo effects. When devices are geographically separated, *inter-destination synchronization* is commonly used to describe the synchronization between their playout processes. It is crucial to provide coherent shared media experiences between remote users, such as multiparty conferencing, Social TV, or MOBA-style online gaming. This sort of synchronization tries to ensure that all users participating in a shared session perceive the same events (i.e., the frames of each media stream are concurrently transmitted to and played on all the receivers) at the same time, regardless of any latency discrepancies between them.

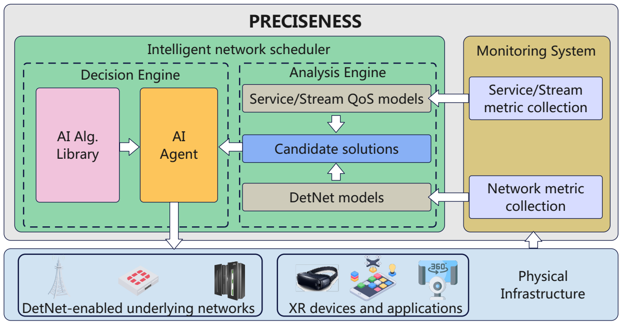
PRECISENESS
Intelligent network scheduler
Decision Engine
AI Alg. Library
AI Agent
Analysis Engine
Service/Stream QoS models
Candidate solutions
DetNet models
Monitoring System
Service/Stream metric collection
Network metric collection
DetNet-enabled underlying networks
XR devices and applications
Physical Infrastructure

Fig. 5. The conceptual management framework of the proposed PRECISENESS framework.

VI. PRECISENESS

To enable the multi-user XR collaboration scenarios and support the XR/Metaverse over the 6G communication networks, the above-mentioned enabling technologies should be synthesized to facilitate the multimedia data transmission and interaction between multi-user in a deterministic way. Thus, we propose the **PRECISENESS**: An AI-em**P**owe**R**ed d**E**terministi**C** mult**I**-u**S**er **E**xte**N**ded r**E**ality re**S**ource orche**S**trator to handle multi-user XR applications for the Metaverse.

The development of PRECISENESS for 6G multi-user XR communications is complicated further by the presence of the following challenges. First, the number of users involved in the multi-user XR/Metaverse is not only featured with a large amount of quantity but also with additional dynamics on the user behaviour due to the user and gaze mobility. Service provisioning for the multi-user should be able to accommodate the large-scale user demand while also taking into consideration the user quality of experience. Supporting a wide variety of XR streams, each with different quality of service criteria, complicates further the design of a service provisioning scheme. Second, ubiquitous intelligence makes it possible for achieving 6G networks. However, facilitating AI capabilities calls for a number of steps to be taken, including the collection of high-quality network and service telemetry samples, proper AI algorithm selection, and the efficient training of AI models and low-latency AI inference that are satisfactory.

Two key components are proposed in **PRECISENESS**: an AI-empowered XR network scheduler and a monitoring system, which can interact with each other to support deterministic stream delivery for multi-user XR by coordinating the service provisioning and network resources. The monitoring system collects the telemetry data of the XR services and network condition in order to learn the characteristics of the XR streams and enables the proactive network configuration and management by predicting the QoE level. It delivers the collected metrics of XR services provided at client sides and the network metrics at network sides to the intelligent network scheduler. After receiving the monitoring data, the Analysis Engine (AE) of the scheduler determines the quality of experience (QoE) model of the XR service, such as the amount of packet loss or MTP delay that XR streams are experiencing at the moment, and the network model, such as the average network bandwidth utilization or the status of network congestion. The input that will be provided to the Decision Engine (DE) will consist of several candidate solutions that have been derived on the basis of the network and service model. Then, an appropriate AI algorithm will be chosen from the library, and utilized by AI agent to determine the optimal network configuration from the candidate solutions, such as the re-routing/reconfiguration decision, the transmission rate adjustment, or the assignment of stream priority.

VII. Case Study

This section demonstrates the effectiveness of PRECISENESS for managing multiple access to the frequency spectrum in a highly dynamic environment for XR UEs. This scenario involves equipment competing for time-slotted channels within a single small cell served by a Small Base Station (SBS). All UEs periodically send their packets using conventional multiple access protocols, such as Time-Division Multiple Access (TDMA). Because of the ever-fluctuating number of active users, the environment is non-stationary. Factors such as user mobility, application switching, or bandwidth-saving strategies contribute to changes in the number of active users in the Metaverse. The term ”context” refers to the set of all active UEs with unique identifiers and their assigned channels. Therefore, context changes occur whenever a UE joins or leaves a channel. It is assumed that XR UEs are informed of the arrivals and departures of other UEs via SBS. However, they are unaware of the transmission patterns, so they must learn to coexist with other UEs.

We will utilize the Continual Learning-Double Deep Q Learning (CL-DDQL) to determine the transmission channel for XR streams. The advantage of using Adaptive AI techniques such as CL-DDQL, the learning agent can remember what it has learned so far, so that when the context changes and

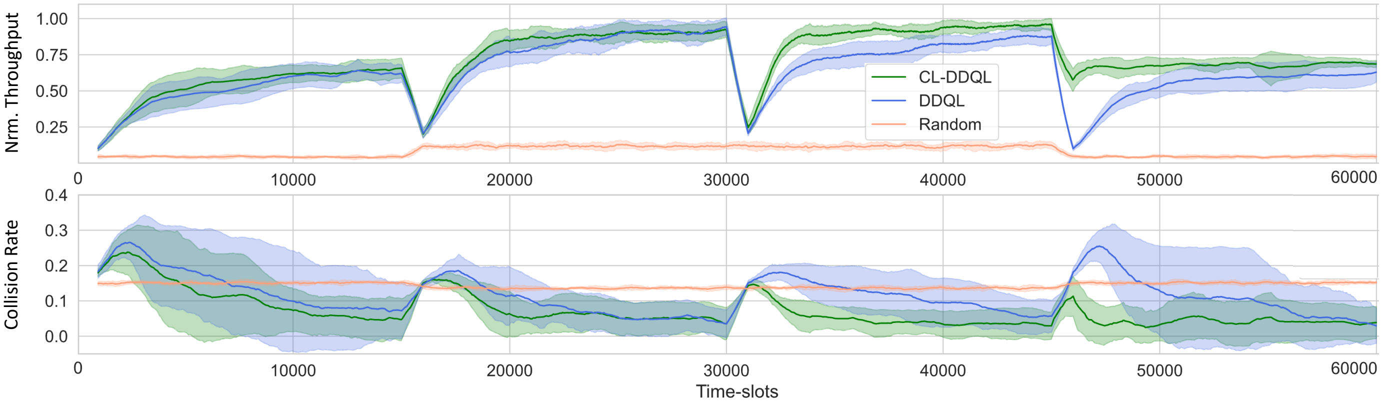

Fig. 6. Normalized throughput and collision rate vs. time slots for CL-DDQL, DDQL, and Random

it encounters a previously learned context, the agent doesn't need to learn it again and can simply start using previously learned weights to make decisions. The core component of CL-DDQL is a Double Deep Q-Learning agent with action space $\mathcal{A} = \{a : (k, c) | k \in \{1, ..., \mathcal{K}\}, c \in \{1, ..., \mathcal{C}\}\}$, where $a : (0, c)$ points to sensing channel c for one time-slot, and $a : (k > 0, c)$ denotes the transmission of a packet with length k on channel c. In this scenario, it is assumed that context transitions occur at specific times, they identifies a TDMA UE that transmits a packet of size k beginning on the t-th time slot of each frame of size f on channel c. Since the agent is designed to maximize its throughput, the reward is equal to the length of successfully transmitted packets. In the case of sensing channel c, the observation set would be $\boldsymbol{O} = \{\textit{Busy, Idle}\}$, whereas it would be $\boldsymbol{O} = \{\textit{Success, Collision}\}$ in the case of packet transmission. The state of the agent is the sequence of the most recent $\mathcal{H}$ (observation, packet length, channel) tuples. Inter-user synchronization will be considered as the constraints where packets from different media streams should be transmitted through the same time slot. To accommodate the non-stationary nature of the environment, this DDQL agent is empowered with the Continual Learning capability, wherein the experience memory and weights of each context viewed are saved and reused when encountering that context again. For example, the first and final quarters of the simulation take place in the same context, the CL-DDQL agent will utilize its prior knowledge of the first context when encountering it again.

The results are illustrated in Fig. 6. There are four contexts, with the first and last contexts being identical. In this scenario, performance comparison is conducted for XR UEs using CL-DDQL, traditional DDQL, and Random methods. The distinction between CL-DDQL and DDQL is that the former features a context management system, whereas the latter lacks memory, so each declared context appears to be new. In the Random method, the XR UE transmits a packet of random length over a random channel. Using the CL-DDQL technique, XR UEs should use their prior knowledge of the original context when re-encountering it. The results verify that CL-based XR UEs have the necessary backward transfer capability for non-stationary environments. Moreover, when confronted with novel contexts, the CL-DDQL XR UEs employ the forward transfer capability. The pre-trained feature extractor enables the CL algorithm to converge substantially faster than the typical DDQL approach, despite the fact that the second and third contexts are dissimilar from the first. In addition, the results demonstrate that DDQL has higher variances in all measures, which is particularly undesirable in quality-sensitive use-cases like the Metaverse. Evidently, Random, the least complex technique, is inefficient as well.

VIII. Conclusion

This article commenced by illustrating the new features of multiple media generation/streaming and its unique challenges in terms of quality of experience guarantee compared with conventional video streaming in multi-user extended reality (XR)/Metaverse. The article then gave an overview on the design of a system along with its essential components to support an immersive and seamless multi-user XR experience. Following this, the article presented a brief analysis into the primary quality of service (QoS) and quality of experience (QoE) demands that XR applications have placed on the underlying networks. Furthermore, the article paid special attention to the XR collaboration scenarios by shedding light on the concept of inter/intra-media synchronization as well as inter-source/destination synchronization, leveraging two potential use cases in multi-user extended reality. As a next step, the article proposed an artificial intelligence-enabled deterministic multi-user extended reality orchestrator as a potential solution to the challenge of providing services to multiple users simultaneously. In the end, the article demonstrated the effectiveness of the proposed approaches by carrying out a continual learning-enabled case study in a single-site, multi-user XR scenario. The findings advocated for the effectiveness of the proposed approach in supporting the delivery of high-quality immersive XR services. In the future, we will investigate on incorporating multiple IoT data streams combined with XR video content to provide multimodal XR service experience with proposed intelligent orchestrator.

Acknowledgments

This research work is partially supported by the European Union's Horizon 2020 Research and Innovation Program through the 6GSandbox project under Grant No. 101096328, the Business Finland 6Bridge 6Core project under Grant No. 8410/31/2022, the Academy of Finland 6G Flagship program under Grant No. 346208, and the Academy of Finland IDEA-MILL project under Grant No. 352428. Prof. Song was supported by the Ministry of Science and ICT, Korea, under the Information Technology Research Center support program (IITP-2023-2021-0-01816) supervised by the Institute for Information & Communications Technology Planning & Evaluation.

References

[1] P. Dogga, S. Chakraborty, S. Mitra, and R. Netravali, "Edge-based transcoding for adaptive live video streaming." in *HotEdge*, 2019.

[2] M. Zink, R. Sitaraman, and K. Nahrstedt, "Scalable 360 video stream delivery: Challenges, solutions, and opportunities," *Proceedings of the IEEE*, vol. 107, no. 4, pp. 639–650, 2019.

[3] H. Yu, T. Taleb, and J. Zhang, "Deep reinforcement learning based deterministic routing and scheduling for mixed-criticality flows," *IEEE Transactions on Industrial Informatics*, 2022.

[4] L. Tian, M. Yang, and S. Wang, "An overview of compute first networking," *International Journal of Web and Grid Services*, vol. 17, no. 2, pp. 81–97, 2021.

[5] S. Kianpisheh and T. Taleb, "A survey on in-network computing: Programmable data plane and technology specific applications," *IEEE Communications Surveys & Tutorials*, 2022.

[6] G. Shi, Y. Xiao, Y. Li, and X. Xie, "From semantic communication to semantic-aware networking: Model, architecture, and open problems," *IEEE Communications Magazine*, vol. 59, no. 8, pp. 44–50, 2021.

[7] K. Khetarpal, M. Riemer, I. Rish, and D. Precup, "Towards Continual Reinforcement Learning: A Review and Perspectives," Nov. 2022, arXiv:2012.13490 [cs]. [Online]. Available: http://arxiv.org/abs/2012.13490

[8] C. D. Alwis, A. Kalla, Q.-V. Pham, P. Kumar, K. Dev, W.-J. Hwang, and M. Liyanage, "Survey on 6G Frontiers: Trends, Applications, Requirements, Technologies and Future Research," *IEEE Open Journal of the Communications Society*, vol. 2, pp. 836–886, 2021, conference Name: IEEE Open Journal of the Communications Society.

[9] T. Taleb, A. Boudi, L. Rosa, L. Cordeiro, T. Theodoropoulos, K. Tserpes, P. Dazzi, A. Protopsaltis, and R. Li, "Towards supporting xr services: Architecture and enablers," *IEEE Internet of Things Journal*, 2022.

[10] T. Taleb, N. Sehad, Z. Nadir, and J. Song, "Vr-based immersive service management in b5g mobile systems: A uav command and control use case," *IEEE Internet of Things Journal*, 2022.

[11] H. Dong and J. S. Lee, "The metaverse from a multimedia communications perspective," *IEEE MultiMedia*, vol. 29, no. 4, pp. 123–127, 2022.

[12] L. Zhang, H. Dong, and A. E. Saddik, "Towards a qoe model to evaluate holographic augmented reality devices," *IEEE MultiMedia*, vol. 26, no. 2, pp. 21–32, 2019.

[13] M. S. Elbamby, C. Perfecto, M. Bennis, and K. Doppler, "Toward low-latency and ultra-reliable virtual reality," *IEEE Network*, vol. 32, no. 2, pp. 78–84, 2018.

[14] T. Scargill, S. Eom, Y. Chen, and M. Gorlatova, "Ambient intelligence for next-generation ar," *arXiv preprint arXiv:2303.12968*, 2023.

[15] Z. Hou, C. She, Y. Li, D. Niyato, M. Dohler, and B. Vucetic, "Intelligent communications for tactile internet in 6g: requirements, technologies, and challenges," *IEEE Communications Magazine*, vol. 59, no. 12, pp. 82–88, 2021.

Hao Yu received the B.S. and Ph.D degree in communication engineering from the Beijing University of Posts and Telecommunications (BUPT), Beijing, China, in 2015 and 2020. He was also a Joint-Supervised Ph.D. Student with the Politecnico di Milano, Milano, Italy. He is currently a Postdoctoral Researcher with the Center of Wireless Communications, Oulu University, Oulu, Finland. His research interests include intelligent edge network, time sensitive networks, 6G deterministic networking.

Masoud Shokrnezhad received his B.Sc. degree in information technology from Shahid Madani University of Azerbaijan, Tabriz, Iran, and his M.Sc. and Ph.D. degrees (as a bright talent) in computer networks from Amirkabr University of Technology (Tehran Polytechnic), Tehran, Iran, in 2011, 2013, and 2019, respectively. He is currently a postdoctoral researcher with the Center of Wireless Communications, University of Oulu, Oulu, Finland.

Tarik Taleb is a Professor at University of Oulu, Finland. Between Oct. 2014 and Dec. 2021, he was a Professor at Aalto University. He also worked as assistant professor at Tohoku University. He holds a B.E. degree in information engineering, and M.Sc. & Ph.D. degrees in information sciences from Tohoku University.

Richard Li received the Ph.D. degree in mathematics with concentration in computer science and artificial intelligence from the Instituto Superior Tecnico, Universidade de Lisboa, Lisbon, Portugal, in 1993.,He is the Chief Scientist and the Vice President of Network Technologies at Futurewei, Santa Clara, CA, USA.

JaeSeung Song is a professor at Sejong University, South Korea. He received a Ph.D. from Imperial College London in the Department of Computing, United Kingdom. He holds B.S. and M.S. degrees in computer science from Sogang University.