IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 30 January 2024, accepted 28 March 2024, date of publication 22 April 2024, date of current version 2 May 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3392509

SURVEY

A Survey on Enabling XR Services in Beyond 5G Mobile Networks

E. STAFIDAS AND F. FOUKALAS, (Senior Member, IEEE)
Department of Informatics and Telecommunications, University of Thessaly, GR35100 Lamia, Greece
Corresponding author: F. Foukalas (foukalas@ieee.org)

ABSTRACT Extended reality (XR) would become the main enabler of future immersive experience services. The challenge today is how to make this new set of immersive experience services running through a mobile network such as beyond 5G efficiently in terms of latency and Quality of experience (QoE). Academics and industry have already started to study and provide solutions towards enabling XR services in beyond 5G networks. These solutions affect both the network and the service elements by either introducing new ones or modifying existing ones. In the present paper, we proffer an extensive survey in this topic with details on the solutions proposed by academia per part of the network, i.e. radio access, core, edge etc. Moreover, we provide a thorough overview of 3rd Generation Partnership Project (3GPP) standards in Releases 17 and 18 focusing on the proposed solutions, Key Performance Indicators (KPIs), services and network elements. Finally, we summarize the lessons gained from this survey and draw the attention regarding open affairs for further research exploration for enabling XR services in beyond 5G networks and towards 6G.

INDEX TERMS Extended reality, immersive experience, 5G-advanced, mobile edge computing, quality of experience.

I. INTRODUCTION

The new digitally connected world of metaverse requires new and complex procedures within existing mobile and fixed internet infrastructures, where the former is considered the 5G network and the latter the edge cloud computing of modern Internet infrastructure. For the fixed networks, where edge and cloud play already an important role in running immersive experience services such as extended reality (XR), in the case of mobile networks such as 5G, the situation is quite more complicated. These holds given the nature of mobile environment that are enabled by wireless medium connectivity that is not always stable. In particular, mobile wireless environment might cause several bottlenecks in running XR services on current and future mobile devices, and this is another interesting topic for research nowadays. To this end, several new elements and solutions are required to make this happen, i.e. enabling XR services in beyond 5G networks, as can be found in the literature and the standardization bodies. Thus, this is the focus of this survey in order to give a comprehensive overview of what has been done so far in this topic and what is the road ahead in terms of open challenges.

More specifically, we present in this survey an overview of the mobile XR services landscape that lists a set of solutions that have been found in the literature. This is considered a state of the art in the field of XR services running on 5G mobile networks. All references are well-structured scientific papers with a view to 5G-Advanced (5G-A) and towards 6G. This first part of our survey is concluded with a taxonomy of the solutions highlighting the part of the network where such a solution has been proposed. In the sequel, the focus of our survey is on the 3rd Generation Partnership Project (3GPP) standardization body, where we first offer a synopsis of the relevant 3GPP standards that deal with XR services applications. Next, an emphasis is given to the 3GPP solutions that can be found in several 3GPP technical documents relevant to Releases 17 and 18. A list of Key Performance Indicators (KPIs) and technical requirements in general follows, which can also be found within the 3GPP spec documents. Finally, the 3GPP architectures of XR services and beyond 5G networks are depicted in two

The associate editor coordinating the review of this manuscript and approving it for publication was Xiaodong Xu.

© 2024 The Authors. This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License.
For more information, see https://creativecommons.org/licenses/by-nc-nd/4.0/

different figures in order to highlight the newly introduced elements for enabling XR services in beyond 5G networks. Our survey concludes with a section where the lessons learned and the open challenges are described.On the one hand, it summarizes what has been achieved so far and, on the other hand what is predicted to be done in the coming days on the subject of XR services in beyond 5G networks. This comprehensive study is considered important for making the metaverse application true in future 6G mobile networks.

In order to make a literature review of relevant works to our survey, we list below a set of most related works that are considered a sort of surveys and overview papers. For example, the work in [1] deals with the outline of 6G communications and networking technologies for the metaverse without analyzing though the actual architecture and the practical needs for enhancements in beyond 5G system. The work focuses mostly on a few challenges such as edge computing, security and digital twin.

The work in [2] presents a layered architecture for the metaverse and how it will be integrated into the future 6G network and it also lists a set of open challenges towards this end. However, their proposal is not associated with the 3GPP standards and in this sense, it does not give a 3GPP based solution but rather a computing architecture. The same holds for the [3] that discusses some general solutions, which are not specific to the standardization and the work has been done within the 3GPP in terms of architecture.

In [4], the authors also present a generic architecture that is relevant to different layers such as physical and virtual worlds. Moreover, they talk about the role of edge computing without referring to the progress within the 3GPP standards and the road ahead. The [5] is more relevant to the XR services and the 3GPP standards for beyond 5G networks and although they list several different architectures for the application layer, the core and the Radio access network (RAN), they miss details about the beyond 5G architecture and the corresponding open challenges. In [6], the authors mention enabling technologies for the future immersive experience; however, they lack of mentioning the overall beyond 5G architecture that will enable future XR services and not only Virtual Reality (VR), where they provided an example.

In [7], one can find a good list of open challenges towards beyond 5G XR services. However, the authors do not deal with the actual 3GPP architecture and the required solutions rather than providing a communication flow at the application layer. There is also a survey in the metaverse in [8] that is not focused on how this will be enabled through the beyond 5G network. However, it is comprehensive and lists many interesting ideas from application and security point of view. Another interesting survey can be found in [9], where the focus is on the VR streaming applications and not networking, where some future 5G directions and solutions are discussed. Further, [10] is a paper on XR services over 5G networks focusing mostly on 3GPP standards evolution and giving also some practical solutions with results. However, they have not extensively provided a comprehensive survey in the area of XR services in mobile 5G networks looking into recent academic results. Most importantly, they do not provide beyond the 5G architectural elements of the network to enable XR services. In particular, they have not presented the overall XR services and beyond 5G network architecture highlighting new elements and solutions including lessons learned and open challenges.

In [11], the authors provide a tutorial on supporting XR services in 3GPP Rel.17 and beyond 5G system talking also about potential solutions in Rel.18. They also provide some results in terms of XR capacity and power. Nevertheless, they do not survey the literature on this topic nor provide a 3GPP based architecture that can enable future XR services since they look into individual components. Finally, in [12], a paper on 5G-A, i.e. 3GPP Rel.17 and 18, the authors devote a section for the XR evolution within 3GPP. They mention several existing specifications and potential new ones in Rel.17 and 18 respectively.

Table 1 summarizes the most relevant surveys mentioned above while provides our contribution that is actually summarized as follows:

a) We provide an extensive survey about XR services in 5G networks and beyond by reviewing the existing literature. This survey is classified into the following categories such as: application layer, networking and edge/cloud computing.
b) We provide a detailed survey regarding the progress of XR services within the 3GPP Rel.17 and 18 standards, listing also the relevant KPIs and traffic attributes.
c) We illustrate an XR enabled 5G-A architecture after careful analysis of XR 3GPP architectures.
d) We summarize both lessons learned and open challenges in this topic.

The rest of this document is categorized in the subsequent sections. Section II provides a survey and taxonomy on XR services in 5G networks based on academic research papers. Section III provides a comprehensive survey in 3GPP XR related works in terms of standards, solutions, KPIs and services and networking architectures. Section IV presents a summary of lessons learned and a list of open challenges and future directions to enable XR services in beyond 5G mobile networks. Sec. V concludes our survey. Fig. 1 illustrates the organization of this survey.

II. THE MOBILE XR SERVICES LANDSCAPE

This section describes the advances in the mobile experience through XR services research topic and details about each solution. At the end of this section, a taxonomy of the different solutions is provided in terms of mobile networking elements. The topics are organized according to the parts of the system that they are considered.

A. APPLICATION LAYER

The solutions below are considered at application layer in collaboration with the networking layer. For example, in [13], the authors address evaluating immersive experiences (IEs)

TABLE 1. Relevant surveys and our contribution.

Reference	Year	Scope
[2]	2023	This work presents a computing layered architecture for the metaverse and how it will be integrated into the future 6G network and lists also a set of open challenges towards this end. The architecture is considered computing on top of the 6G network without integration with the 3GPP elements.
[7]	2022	This paper provides a good list of open challenges for beyond 5G XR services, including requirements. The authors do not deal with the actual 3GPP architecture but rather provide a communication flow at the application layer.
[10]	2023	This paper focuses on XR services over 5G networks, considering 3GPP standards evolution and gives some practical solutions with results. They give an overview of XR services in 3GPP while providing details about specific KPIs and also an evaluation of XR services in 5G New Radio (NR) Rel.17. They have not presented an overall XR services and beyond 5G network architecture highlighting new elements and solutions.
[11]	2022	The focus of this paper is on assessing the capability of XR services in 3GPP Rel.17 beyond 5G system. The study examines three KPIs such as XR capacity, power-saving techniques for user equipment (UE) and improved UE orientation and positioning through network localization. The paper proposes several areas of investigation within the 5G domain, including the implementation of more dynamic radio resource allocations and the integration of ML techniques and Artificial Intelligence (AI) to be adjusted to dynamic XR traffic more effectively.
[12]	2023	This work focuses on 5G-A specifications towards 6G and provides a short section about the XR evolution within 3GPP. It provides a description of the service requirements and key considerations for the support of XR in 3GPP, talking about specific standards.
Now	Ours	Our work first provides an extensive survey about XR services in 5G networks and beyond by reviewing the existing literature. This survey is classified into the following categories such as application layer, networking and edge/cloud computing. Next, it provides a detailed survey regarding the progress of XR services within the 3GPP Rel.17 and 18 standards, listing also the relevant KPIs and traffic attributes. Moreover, we illustrate an XR enabled 5G-A architecture after careful analysis of XR 3GPP architectures. Finally, a list of both lessons learned and open challenges conclude this paper.

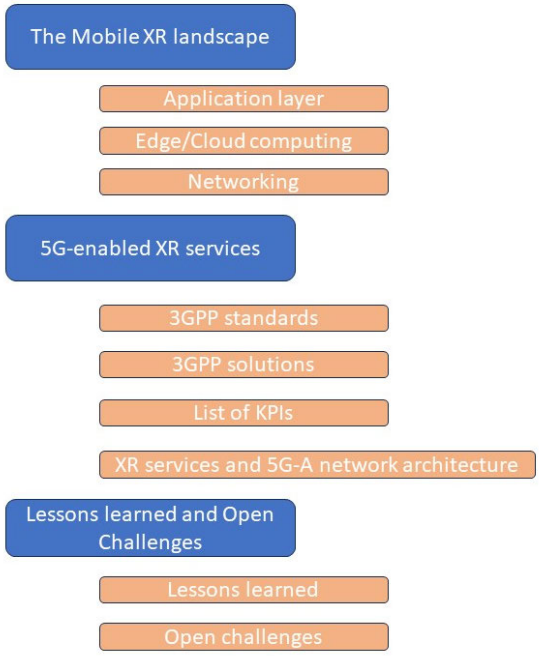

FIGURE 1. Organization structure of this paper.

in network services. They propose AI-based methods like generating multi-view signals and mathematical analysis. These methods outperform traditional models for different data scenarios. The paper also suggests exploring online IE evaluation for responsive network services.

In [14], authors leverage blockchain for secure Augmented Reality (AR) and VR in 6G networks. They introduce a decentralized architecture, analyzing latency and packet processing. The paper lays the foundation for trusted decentralized AR and VR in 6G and explores their potential for massive Internet of Things devices (mIoT) applications.

In paper [15], authors suggest leveraging blockchains and Information Centric Network (ICN) for improved VR/AR experiences in 6G-enabled massive IoT. They introduce Proof of-Cache-Offloading (PoCO) and a permissioned blockchain for secure content and resource transactions, enhancing Quality of Service (QoS). The paper's proposed cache index (CI) algorithm and blockchain-enabled cache model for VR/AR content receive validation through simulations, showcasing their potential for 6G networks and edge infrastructure.

Paper [16] introduces an XR system for collaborative cross-platform applications using Unity and Mirror, facilitating synchronized experiences. It streamlines platform-specific components and allows multi-platform app development without configuration changes. The paper includes a collaborative game example and provides guidance for future applications using existing standards and open-source tools.

In [17], an innovative AR-based support system is deployed at the Vodafone Transceiver Base Station (BS). It enhances worker safety in hazardous environments using advanced sensors and 5G technology. The system employs hazard-detection algorithms and a holographic viewer, consisting of an Augmented Reality (AR) helmet, Wireless Local Area Network (WLAN) connectivity, customer-premises equipment (CPE) for 5G connection and real-time processing via a Multi-access Edge Computer (MEC). Future plans involve immersive integration of augmented data with the real environment.

In [9], a comprehensive survey investigates adaptive streaming for 360° video content, particularly in AR and VR applications. It covers content creation, formats, delivery and quality adaptation. Topics include rate adaptation, segment scheduling, network-aware approaches, quality assessment, viewport-dependent streaming, bitrate adaptation, bandwidth estimation, latency and packet loss. The paper outlines future research areas such as cloud-based streaming, compression, security and Quality of experience (QoE) enhancements.

In [8], an analytical survey explores metaverse privacy and security. It covers threat vectors like identity theft, access control and malware attacks. The paper introduces a decentralized metaverse framework, emphasizing user autonomy and secure data management. It discusses AI, honeypots and Software Defined Network (SDN) for situational awareness, cyber-insurance and governance trends. Blockchain is proposed for trust-free digital identities, data reliability, virtual economies and self-governance in metaverse communities.

B. EDGE/CLOUD COMPUTING

The following solutions are considered for the telco edge and cloud computing. For example, in [18], the authors address Cloud XR streaming QoE. They present a mathematical model measuring QoE based on factors like bitrate and interruptions. An optimal bitrate adjustment method is introduced. Simulations validate the model's accuracy and the effectiveness of bitrate adjustments for higher average bitrates while limiting stalls. Future work includes complex traffic patterns and interference from other video streams.

In [19], authors optimize QoS for context-aware AR (ConAR) using the ConAR system on edge and cloud infrastructures. 5G is ideal for low-latency tasks, while the cloud serves as backup. Data transfer delays favor 5G over LTE. Edge exhibits higher jitter, but it's the primary platform for situation assessment models. Future plans include cooperative AR experiments and performance analysis.

In [20], a web-based AR solution with a focus on Mobile Edge Computing (MEC) is discussed. Two real-world use cases are presented. An edge server optimizes image matching using the Affine-Scale Invariant Feature Transform (ASIFT) algorithm, improving response time, fps, feature point matching and terminal power consumption. Three implementations are compared, with edge computing proving the most efficient. The paper suggests future research in optimizing network resources, improving app performance and reducing latency.

In paper [21], a MEC platform is introduced for AR services, leveraging MEC, 5G and AR for high-speed, low-latency performance. This platform serves as a foundation for AR app development, enhancing efficiency and the user experience. Two use cases, Face Recognition and Navigation, demonstrate its feasibility and effectiveness. A numerical evaluation compared it to cloud computing, highlighting MEC's advantages in performance, network reliability and data privacy/security for AR services.

Paper [22] explores edge computing with caching techniques for advancing 5G and beyond wireless networks. It reduces duplicity and redundant traffic with caching at both the core and edge. Computation tasks can occur at either the edge server or the cloud depending on their complexity. The focus is on improving QoS and energy efficiency, emphasizing flexibility in future wireless networks. SDN enhances computing and caching in mobile networks. Future progress will support AR, VR and autonomous systems, where edge caching and computing ensure consistent performance across applications.

In paper [23], MEC and 5G are explored for enhancing Mobile Augmented Reality (MAR) and VR experiences by reducing application delay and bringing computation closer to users. The paper introduces computation handoff(Comp-HO), an algorithm optimizing signal strength and computational load for MEC handoff in 5G. While it adds some overhead, it reduces end-to-end delay compared to benchmark handoff algorithms in MEC scenarios. The paper aims to improve Comp-HO and evaluation methods in the future and develop algorithms addressing interference types and handoff rates in 5G networks.

In paper [24], strategies for minimizing inter-player latency in wireless multiplayer VR games are discussed. It proposes using edge computing for real-time rendering and introduces an algorithm breaking down the non-convex topic into convex sub-topics for iterative optimization. This algorithm considers constraints like frame per second (FPS) and prediction using prerendered field of view (FOV) images. Numerical results show that it reduces inter-player latency compared to a baseline, improving gaming experiences. The study also explores the impact of channel quality on edge resource allocation and content sizes.

In [25], the authors discuss a fly-edge transcoding strategy for reducing motion-to-photon delay in VR videos by compressing frames within the assigned bandwidth, regardless of content. It works with network slicing, ensuring consistent frame rates and dynamic resource allocation based on predicted capacity. Simulations confirm its efficacy for low-delay services across network scenarios, with a picture quality trade-off. Success depends on edge node computational capacity. Future plans include realistic MEC models via BS and slicing operations, optimizing resources and maintaining low latency, with picture quality linked to BS computational capacity.

In [26], the research enhances video-driven AI tasks in a multi-user MEC system by minimizing energy consumption, overall delay and improving accuracy. It uses computational complexity and accuracy models to estimate computation delay and its relationship with input frames. Two algorithms (Search-based and GP-based) optimize resource allocation for Deep Neural Network (DNN) inference tasks, whether local or offloaded to an edge server. Offloading decisions are optimized using the Distributed Alternating Direction Method of Multipliers (ADMM) and a Channel-Aware heuristic algorithm. The paper demonstrates the algorithm's effectiveness in enhancing video-based AI tasks in a multi-terminal MEC environment through simulations and experiments.

In paper [27], a detached learning approach enhances live VR video streaming over wireless networks. It uses MEC association and FOV prediction. A Recurrent Neural Network (RNN) model based on Gated Recurrent Unit (GRU)

architecture forecasts individual VR user's FoV preferences over time. Distributed Deep Reinforcement Learning (DRL) and Centralized methods establish associations between FoV demands and user locations, optimizing VR user group-MEC correlations and rendering MEC selection for relocation. This improves long-term QoE for VR users. Simulations show that this MEC rendering, combined with DRL and RNN algorithms, significantly enhances QoE and reduces VR interaction delay compared to rendering solely on VR devices.

In paper [28], the focus is on minimizing inter-player latency in wireless multiplayer VR games through effective bandwidth and resource allocation using MEC. An innovative framework examines player interactions and post-processing at either the mobile device or MEC server. The non-convex problem, considering pre-rendered FOV images, is efficiently solved using the Nonconvex Primal-Dual Splitting Method for Distributed and Stochastic Optimization (NESTT-G) algorithm. Results show that this algorithm significantly reduces inter-player latency with lower complexity and faster convergence compared to standard methods.

In [29], the authors evaluate the QoE of the AR app ''ViewToo Face Mask Guide'' on 4G LTE, HSDPA+, and 5G networks. They find that 4G LTE and HSDPA+ improved app load times but had lower FPS in HSDPA+ due to data rate and latency. Upgrading to 4G LTE-A reduced the delay. On 5G, Sub-6 GHz improved data rates and latency, exceeding 1 Gbps in mmWave 5G, doubling app load speed. Overall, 5G met QoE requirements for AR, reducing load times. The paper suggests future research should explore complex app impact on 5G QoE.

In [30], the paper explores how the changing cloud and network architecture affects AR and VR. They use a methodology to assess QoE for VR and AR across 5G and LTE. They break down AR/VR operations to identify key quality indicators (KQIs) that impact user satisfaction. Two QoE models are created using network operator performance data. These models correlate application KQIs with network KPIs using operator RAN KPIs and app performance data. Results show that LTE KPIs are inadequate for satisfactory QoE in AR and VR apps. They also evaluate a 5G Non-Standalone (NSA) performance assumption for VR gaming in a future network model with increased capacity.

In [6], the paper suggests enhancing QoS for multisensory XR-aided devices. They use a VR teleoperation testbed and a platform for remote-controlled unmanned aerial vehicles (UAVs) to evaluate service quality. The study shows that end-to-end (E2E) command and control delay remains consistent across heights and frequencies but surges at 60 Hz due to buffer overflow. Video streaming delay is more noticeable for flying UAVs due to limited antenna gain from BS sidelobes. The authors conclude that current wireless networks don't meet the latency needs for haptic robots and UAV teleoperation, suggesting MEC as a potential solution. Future developments and tailored solutions are essential to address these challenges effectively.

This paper [31] enhances wireless VR QoE in 5G by using QoE-aware transmission, buffer-aware caching and data correlation. They coordinate resource allocation using a quantum-inspired reinforcement learning (QRL) algorithm and evaluate it with 15 distributed virtual equipment (VEs) across 4 intelligence-assisted access points (IAPs). The approach reduces VEs' energy consumption through computation offloading and achieves a near-optimal system reward. The study concludes that a comprehensive system strategy and algorithm design can make wireless VR widely accessible, emphasizing the need to address security concerns in the future.

In paper [32], the authors investigate challenges in globally implementing immersive services, focusing on remote-controlled virtual reality for robots. They analyze factors like display refresh rate, streaming frame rate and technology's impact on Glass-to-Glass latency. Remote devices have 360-degree cameras and are controlled via hand motions and Head-mounted display (HMD) controllers, communicating using real-time protocols. Latency is affected by camera refresh and streaming rates, which can be reduced by increasing the latter but require substantial computational power. The paper concludes that leveraging technologies like 5G, AI and cloud edge computing shows promise for immersive deployments but highlights the need for further research to enable global-scale immersive services.

In paper [33], authors propose a collaborative approach for object identification in 5G mobile Web AR, focusing on edge-based implementation. They evaluate three DNN service provisioning techniques (Shortest Job First (SJF), First Come First Serve (FCFS) and Highest Value First (HVF)) for AR-based instance retrieval and recommendation. Their edge-assisted collaborative method aims to balance provider and user interests, considering metrics like balance rate and mobile energy consumption. The authors suggest heterogeneous computing collaboration to improve DNN inference delay and energy consumption in mobile Web AR. They also recommend using neural architecture search for more efficient DNNs. Results indicate that collaborative methodology surpasses edge-based and autonomous cloud offloading for AR apps in 5G networks.

In [34] Edge AR X5 framework is introduced for Multi-User Collaborative Mobile Web AR in 5G networks. It includes the Panda Betrayal mobile AR app accessible via URL. The framework employs Motion Aware Run-time Scheduler (Mo-KFP) and BAS-Based Communication Planning (BA-CPP) mechanisms for collaborative communication and computing. BA-CPP balances internet service provider and user requirements, while Mo-KFP facilitates collaboration through Device-to-Device (D2D) communication, reducing initialization times. Experiments show lower communication delay and mobile energy consumption compared to D2D alone. The paper also discusses potential enhancements for mobile Web AR, such as On-Web AI Service, multi-edge collaboration AI, 5G-Enabled Networking AI and network slicing.

In paper [35], a precise elastic calculation partitioning approach for DNNs in 5G networks is presented. This approach uses D2D and MEC technologies to optimize resource coordination. The paper includes a use case involving a recommendation application and AR-based resource retrieval for mobile web advertising. It introduces estimation models for per-layer inference delay and energy consumption in DNNs and a versatile DNN partitioning technique. Experiments conducted on a 5G trial network with a web-based mobile AR app demonstrate the effectiveness of this cooperative approach in meeting application performance requirements while minimizing implementation overhead. This approach supports precise computation partitioning through layer granularity offloading, enabling decentralized collaboration and adaptive computation scheduling.

C. NETWORKING

The following set of solutions are considered as networking ones. In particular [36] addresses diverse network requirements for XR technologies. It defines various XR flavors to emphasize the need for high-reliability networks with Mbps-level bitrates and 10-20 ms latencies. Meeting XR's demands for latency, reliability and bitrates requires robust link adaptation, delay-aware scheduling and normalized features. Simulations show that basic XR flavors are generally supported in current networks, while denser networks support more complex options. Less densely deployed networks may need upgrades for coverage. AR apps can effectively use existing 5G tech with lightweight glasses and three-dimesion (3D) compression. Edge computing enhances 3D streaming and reduces energy consumption and latency. Incorporating new media formats like point clouds and offloading XR processing to the mobile network edge can improve XR communication scenarios while addressing device limitations.

In [11], the focus is on evaluating XR services (Release 17) in 5G NR. The study assesses three key aspects: XR capacity, UE power-saving techniques with enhanced UE orientation and positioning via network localization. Results show non-linear scaling of XR performance with data rate and significant impact of delay. Four power-saving techniques are analyzed, with the ''On duration'' enhancement reducing power-saving benefits. Notably, the AR/VR setup at 45 Mbit/s poses challenges for power-saving. The paper suggests further research areas in 5G, including dynamic radio resource allocation and AI integration for dynamic XR traffic. In summary, 5G NR supports XR services, but improvements are needed for higher data rates, reliability and reduced interruption during handovers.

The authors in [37] present a methodology that uses a software suite to manage and control various components, including gNodeB deployment, Remote Radio Unit(RRU), Baseband Unit (BBU) and UE radio. The main goal is to enhance and design a gNodeB using MIRACLE radio technology and APIs. They introduce a RAN Intelligent Controller (RIC) software system to support RAN resource optimization adaptable to different objective functions (Prioritization Protocols). They also incorporate MEC for AR/VR integration into the user platform. A use case involves slicing for AR, VR and XR experimentation and RAN-aware content enhancement to improve user QoE. The methodology includes 3D scene generation at the mobile edge for Multi-Mission Headsets (MMH) to measure uplink and downlink efficiency. Low Latency Mobile Edge Computing (LLMEC) runtime dynamically configures routing to minimize delay for each user's 3D scene generation application.

An innovative approach in [38] for VR and AR in 5G cloud-based networks is introduced. It uses SDN to reduce network latency, employs a multi-path cooperative route (MCR) for efficient wireless transmissions from edge data centers (EDCs) to end-users while minimizing energy use. The authors present a Service Effective Energy (SEE) framework to assess MCR's energy consumption in VR and AR deployments and introduce a Service Effective Energy Optimization (SEEM) algorithm to reduce SEE in 5G microcell networks. Simulations show that the MCR model outperforms traditional single-path routing for VR and AR deployments. SEEM is tailored to reduce network energy while maintaining latency below a set threshold, making it more suitable for VR apps compared to AR apps.

In paper [39], a method is introduced for delivering VR and AR experiences through multicast rather than individual transmission. It explores using LTE and 5G networks for widespread VR data distribution, with simulations assessing its effectiveness under different scenarios, considering user placement. Results show that VR broadcast can be significantly improved, especially with 5G millimeter-wave microcell networks, despite slightly increased bandwidth requirements. The study identifies limitations in mobility and optimal positioning of gNodeBs (gNBs) and suggests further investigation for bandwidths exceeding 1 GHz. The paper emphasizes the need for advancements in supporting both individual and multicast streams, implementing beamforming for multicast and balancing spectral efficiency with BS costs.

A comprehensive assessment of XR deployments on a 5G NR infrastructure operating in sub-6 GHz (FR1) and millimeter-wave (FR2) frequency bands is presented in [40]. A split-rendering framework is used to optimize the user experience while keeping device power consumption acceptable. This framework delegates rendering and encoding computations to an edge server via a 5G connection. Simulations consider a 5G cellular system with multiple users and Base Stations (BSs) randomly distributed in their coverage area. Results show that 5G performs well in both FR1 and FR2 scenarios for XR deployments. The paper also discusses potential 5G enhancements to further improve the user experience, including traffic-awareness, semi-persistent scheduling, interference coordination, network coding and mobility-oriented improvements.

Two novel power-saving techniques for extended reality services, which have stringent quality requirements not addressed by existing methods, are studied in [41]. One technique synchronizes discontinuous reception (DRX) with video frame arrival and the other reduces unnecessary physical downlink control channel (PDCCH) overhead. Simulation results show significant power savings with minimal capacity loss compared to existing DRX methods. The paper also discusses potential enhancements, such as extending PDCCH monitoring duration to address packet jitter and suggests further research on power-saving improvements like aligning downlink and uplink for User Equipment assistance information transmission.

In [42], paper explores power-saving strategies for XR applications in 5G NR. These strategies address the challenges of high data rates, low-latency XR traffic and limited battery capacity. The discussed techniques include Connected mode discontinuous reception (CDRX), Low power BWP (Bandwidth Part), Cross-slot scheduling, Scell dormancy and enhancements introduced in Release 17 and beyond. Meticulous simulations, using a 3GPP-specified UE power consumption framework, show that these techniques can reduce UE power consumption up to 40%. The evaluation considers various transmission/reception power states and models frame-based traffic arrival and slot-based power.

Reference [43] addresses the need for power reduction in XR devices in 5G. They propose an Adaptive DRX (ADRX) scheme to flexibly adjust power usage while maintaining QoS. A split-rendering architecture is used to reduce power consumption. The approach achieves up to 10% power reduction compared to constantly active users. In Dense Urban scenarios where standard DRX fails, ADRX supports VR and AR services with up to 4 user devices per cell in 5G-A systems.

In [44], the paper addresses XR streaming challenges and introduces an open-source traffic simulation framework in Network Simulator 3 (ns-3). This framework generates synthetic traffic traces for protocol testing and network improvement. The paper evaluates the model through simulation campaigns, highlighting results and limitations. It also presents a use case with multiple users sharing network resources. Future work includes exploring scheduling techniques for XR traffic alongside other applications and validating the model for higher FPS values. This model is a valuable resource for researchers enhancing XR traffic support.

In [45], it is introduced a novel wireless XR framework in Beyond 5G (B5G) systems with the Internet of Intelligence. This architecture enables adaptable spectrum utilization and seamless handover of XR users across multiple carriers, enhancing user credibility at the network edge while minimizing latency. The paper evaluates user experience in Carrier Aggregation (CA) and flexible spectrum access (FSA) architectures, examining network capacity for extended reality applications. Simulation results show that FSA outperforms CA, with a nearly 20% performance improvement in multi-user scenarios. FSA supports around 50% more user devices in XR services, ensuring legibility and meeting XR application delay requirements.

References [46] and [47], propose two novel algorithms, referred to as improved Outer Loop Link Adaptation (eOLLA), for CBG-based transmissions to enhance downlink system capacity for XR traffic. These algorithms optimize CBG-based transmissions by monitoring desired inaccuracy rates using Hybrid Automatic Repeat reQuest (HARQ) feedback. They achieve a significant system capacity gain, with a 33% and 67% improvement compared to conventional OLLA with a 10% Transport Block Error Rate (TBER) target. The second algorithm effectively manages CBG inaccuracy during the second transmission. The goal of eOLLA is to reduce radio resource usage for XR traffic, accommodating more satisfied XR users. The papers also introduce an enhanced Channel Quality Indicator (eCQI) method for 5G-A XR use cases, providing feedback to gNB for improved link adaptation and Code Block Group(CBG) error rate control, resulting in gains ranging from 17% to 33% compared to the baseline CQI approach.

In [48], the paper explores two XR offloading scenarios using real-world traffic data and models. The first scenario involves transferring sensor data from XR devices to a nearby server or MEC for immersive XR frame display, while the second focuses on delegating heavy Machine Learning (ML) algorithms. They use Johnson's SU distribution to model various parameters and evaluate synthetic data with a 5G RAN emulator. The paper suggests that these models and data are valuable for improving wireless network solutions in both industry and academia, offering an alternative approach for XR offloading scenarios. It also emphasizes the significance of estimating inter-packet time intervals for resource allocation methods.

Reference [49] introduces a solution for establishing a 5G-A system architecture to support XR services. It discusses current advancements in the 3GPP standard, focusing on network architecture enhancements for XR methods. The paper highlights challenges in providing real-time interactive XR services, including lossless audio compression, multimodal synchronization and high-resolution video transmission. It also explores QoS improvements for power-saving and XR transmission. While noting these advancements, the paper acknowledges future challenges, particularly in the context of Metaverse services. These challenges involve integrating architecture with communication, computing and sensing, as well as enhancing IE for emerging media formats.

A novel framework for adaptive quality 360-degree virtual reality streaming is introduced [50], utilizing multicast through sidelink communication. This approach enables the BS to simultaneously transmit video streams to multiple users. The paper considers two scenarios: independent decoding and joint decoding. In overloaded situations with bandwidth and tile quality constraints, the system ensures a

high-quality user experience. An algorithm is presented to optimize system performance by managing sidelink sender selection, bandwidth allocation and tile quality levels for efficiency across all users. The algorithm employs two-stage iterative techniques with low computational complexity, including greedy exploration and progressive relaxation. Extensive simulations show its superiority over baseline schemes, significantly improving utility efficiency and user QoE compared to traditional multicast methods and baseline approaches.

Reference [51] proposes a solution for uplink wireless VR networks by combining offline and online learning algorithms with preemptive retransmission. The offline learning algorithm employs Gated Recurrent Units (GRUs), Neural Networks (NN), K Cross Validation for n-order Linear Regression (LR) and Long Short-Term Memory (LSTM) for each video, collectively trained for all VR videos to forecast user's viewpoints in uninterrupted time slots. Online learning algorithms continuously update their parameters based on current viewpoints from new end-users via uplink transmission to improve prediction accuracy. The preemptive retransmission model enhances uplink transmission reliability and refreshes input viewpoints and online learning parameters. Simulations demonstrate the effectiveness of real-time GRU algorithms with preemptive retransmission in achieving optimal forecasting accuracy.

A multi-client MAC scheduling approach [52] is introduced for wireless virtual reality services in a 5G Orthogonal Frequency Division Multiplexing (OFDM) and multiple-input, multiple-output (MIMO) system. The approach aims to optimize concurrent VR clients while ensuring ultra-high responsiveness, transfer reliability and data rates. It includes three key functionalities: weight calculation based on frame delay and differentiation, dynamic block-error-rate (BLER) targeting for link adaptation and spatial-frequency end-user selection using maximum aggregate delay-capacity utility (ADCU). The paper also presents a simplified algorithm for downlink MIMO recipient selection. Simulation results show a 31.6% increase in the maximum number of served virtual reality users simultaneously compared to conventional BLER-based link adaptation and maximum-sum-capacity scheduling.

Reference [53] proposes a real-time XR loopback mechanism to adapt XR traffic based on 5G network feedback. It introduces and implements various XR loopback algorithms in ns3 simulations. It also studies the validation and impact for various loopback configurations, strategies and algorithms, specifically in an authentic 5G environment featuring numerous XR end-users and a blend of AR/VR/CG and Voice Over IP (VoIP) data transmissions. The results demonstrate that the XR loopback mechanism effectively controls QoS requisites, like data drop, while leveraging favorable channel conditions and elevating the codec rate. The study highlights the significance of XR traffic shaping and the need to configure adaptation algorithms and parameters based on traffic characteristics. The research sets the foundation for further investigations into XR loopback mechanisms for improved QoS in beyond 5G networks.

D. LANDSCAPE SYNOPSIS

The mobile XR services survey presented above in this section concludes to the fact that there are three main studies in the last couple of years related to enabling XR services in 5G and beyond 5G networks. The three categories are depicted in the Fig. 2 below in contrast to an existing/conventional 5G mobile network and its main elements such as RAN, edge, application and Internet. In detail, Fig. 2 depicts the need for adaptive XR streaming that is running in the upper layers of the network architecture, the need for a context - aware edge computing deployed through layers 4 and 5 while a QoS/QoE management is taken place in a cross - layer design fashion across layers 1, 2 and up to 4 and 5. The adaptive 360° streaming is essential for running high quality mobile XR applications. The context - aware edge computing will enable the immersive experience to be associated with the surrounding environment/place. Finally, the CLD on RAN with the transport layers is necessary to provide QoS/QoE management taken into account the wireless medium. It is evident that these three types of solutions are important and it turns out that they drive the current standardization efforts as can be figured out from our relevant survey in the section below.

III. 5G-ENABLED XR SERVICES

This section provides a discussion about the relevant 3GPP standards, solutions and list of KPIs in regards to the 5G enabled XR services. The section concludes with the 3GPP XR services architectures and the beyond 5G network architecture, which depict all functional elements needed to run XR services in beyond 5G networks.

A. 3GPP STANDARDS

In this 3GPP Technical Specification (TS) [54] they are referred the types of positional tracking of XR scene. These are outside-in tracking, inside-out tracking, world tracking and Simultaneous Localization and Mapping (SLAM). As regards spatial mapping the methods are spatial anchors, SLAM and visual localization. To integrate XR applications within the 5G it is proposed an approach that fits the approach of 3GPP Media Streaming in 5G.

In this architecture, there are the following functions: A 5G-XR client is designed to receive 5G-XR session data on a UE and provides interfaces, or Application programming interfaces (APIs) for access by a 5G-XR Aware Application. This client comprises two core sub-functions: the XR Session Handler, which is responsible for establishing and controlling the delivery of XR sessions in coordination with the 5G-XR AF and offering APIs for the 5G-XR Aware Application to utilize; and the second sub-function, the XR Engine, which facilitates communication with the 5G-XR Application Server (AS) to access XR-related data and functionalities, including data processing, sensors,

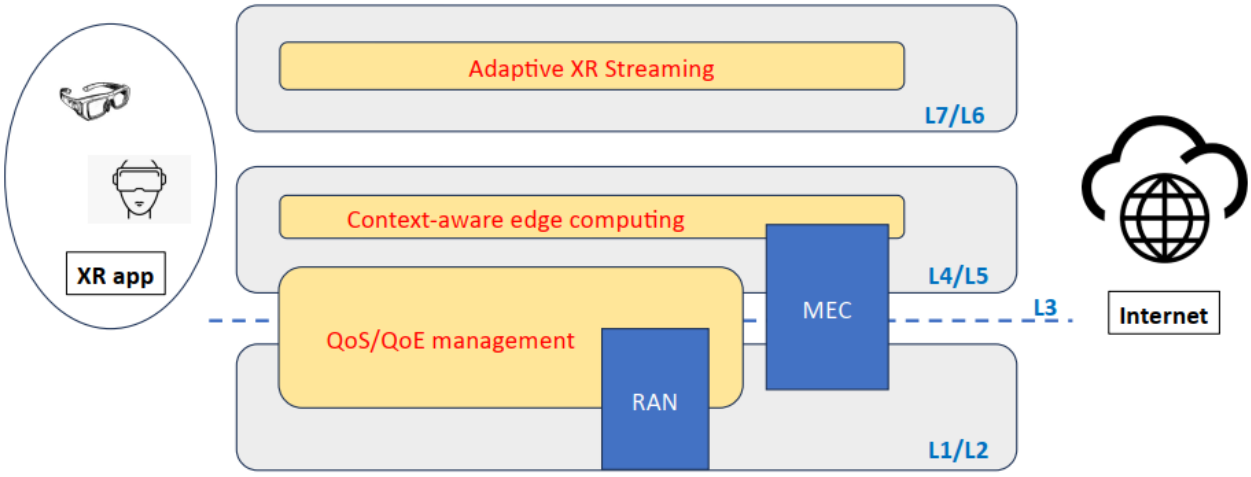

FIGURE 2. Type of solutions to enable mobile XR services in 5G networks.

tracking and interaction with the XR Session Handler for XR session control. Conversely, a 5G-XR Aware Application leverages the capabilities of the 5G-XR Client and network functions (NFs) through provided interfaces and APIs. Typically, the external XR Aware Application controls the 5G-XR Client, implementing logic specific to the application service provider and establishing an XR session. Lastly, a 5G-XR AS serves as the host for 5G-XR media and relevant functions within its AS role.

A 5G-XR Application Provider serves as an independent XR application provider, utilizing 5G-XR client and network operations to deliver an immersive XR encounter to 5G-XR Aware applications. The 5G-XR Application Function (AF) offers diverse control tasks to the XR Session Handler on the UE as well as the 5G-XR Application Provider. It can trigger requests for specific Policy or Charging Function (PCF) treatments and communicate with diverse network operations. The 5G QoS model accommodates QoS flows with and without guaranteed bit rates, supporting Introspective QoS. A QoS Flow ID (QFI) is used to identify flows, which can be dynamically assigned or match the 5G QoS Identifier (5QI). User plane traffic allocated to a corresponding QoS Flow within a Packet Data Unit (PDU) Session receives an identical traffic forwarding approach. 5G-XR applications can leverage both ASs and edge processing for enhanced performance. 3GPP has defined a comprehensive group of application layer connections targeted for Edge Computing that enable discovery, exposure and management of Edge Applications, enhancing the performance and capabilities of 5G-XR applications [54].

In 5G-A era, efforts to enhance edge computing will continue with the goal of bringing the AS closer to the UE. Edge computing is important for enabling XR and Cloud Gaming (CG) and can assist with processing on power and resource constrained devices. The UE sends sensor data to the cloud for rendering, where NR and network slicing can be useful for managing diverse XR services through the 5G network [54]. In the same TR, the authors also developed the following XR Processing and Media Centric architectures. Viewport-Independent delivery, in which sensor information and tracking is processed in XR device and the whole XR scene is received and decoded accordingly. In the Viewport-dependent Streaming methodology, tracking information is handled within the XR device, while the XR delivery engine receives real-time pose data to include it in responsive media requests. Viewport-independent delivery architecture is a method of implementing experiences that can adapt to different display sizes, aspect ratios and FOV configurations, while a Viewport-dependent architecture delivers only the portion of that experience that is visible within the user's viewport [54].

Viewport Rendering in Network refers to a method of enhancing the delivery of XR experiences over a network by prioritizing the rendering of content within the user's viewport. Media decoders are responsible for decoding the media content within the XR device, while the viewport is rendered directly without relying on viewport information [54]. In Raster-based split rendering, an XR server hosts an XR engine that generates the XR scene using data received from an XR device. This corresponds to a method for improving virtual reality performance by rendering separately different parts of a scene for each eye and adjusting the viewpoint to fit user's head movement. The server conducts viewport rasterization and pre-rendering, while the device asynchronously corrects the pose to accommodate pose variations.

In Generalized XR Split Rendering, XR server pre-renders the 3D environment into a more streamlined format for processing by the device. The device subsequently retrieves the pre-rendered media and performs final rendering, applying local adjustments based on current pose [54]. It divides the rendering workload between multiple processors or devices to ameliorate performance and reduce latency. Also, XR Distributed Computing in where the XR processing workload is divided between the device and the XR server [54]. It refers to the use of network devices to execute XR applications, by enabling more complex and intensive resource experiences that could be elaborated on a single device.

XR Conversational Architecture refers to IE that incorporates VR and AR experiences with natural language processing to build engaging communication with users. The discussed services extend the use of Multimedia Telephony Service for IMS (MTSI) and leverage IP Multimedia Subsystem (IMS) for session signaling. In this architecture, a new interface is introduced for media handling within the network. Equivalent to the Service Resource Function (SRF), this interface is expected to support different types of media processing extensively. To enable XR conversational services, signaling extensions are required to incorporate VR/AR specific features and media and metadata frameworks need to maintain appropriate profiles, metadata and codecs. XR Conferencing architecture declares technical infrastructures and platforms that allow geographically scattered users to interact in a shared environment by exploiting VR and AR technologies. Finally, CG refers to the use of cloud computing in gaming experiences that contain VR, AR or Mixed Reality (MR) elements [54].

In this TR [55], they are presented some XR enhancements regarding NR. Awareness in both uplink and downlink conduces to improvements of gNB radio resource scheduling based on the notions of PDU set and Data Burst. More specifically, the creation of new BS tables to reduce errors in Buffer Status Report and the delay in knowledge of buffered data. There are proposed some alternatives between the mapping of PDU sets and QoS flows in the Non-access stratum (NAS) and also regarding the allocation of QoS flows to Data Radio Bearers. Moreover, cross-layer XR QoS optimizations have demonstrated their ability to enhance XR efficiency in 5G and beyond networks. By utilizing an XR adaptation mechanism, the XR experience can be improved based on the specific network conditions. Additionally, if the 5G-A RAN has knowledge about the XR application flows, it can optimize the scheduling and prioritization of XR traffic in the RAN to ensure reliable XR QoS.

Furthermore, the results indicate the need to employ advanced methodologies such as AI/ML algorithms for the coordinated configuration of XR-loopback variables. This approach ensures optimal performance by adapting to the specific attributes of both the 5G channel and network statements and XR traffic. By incorporating insights from cross-layer XR QoS optimization, AI/ML algorithms can leverage information from different tiers of the 5G protocol stack and the XR application ecosystem to make well-informed decisions and enhance the overall XR experience.

In this TR [56] it is presented an E2E XR Traffic architecture and system model that includes interfaces supported by traces for simpler system modeling and simulation. The traces could be v-traces (video traces), s-traces (slices known from H.264/AVC and H.265/HEVC application data units as outputs from a video codec), p-traces (that provide a time series of IP packets associated with different application flows and/or QoS Flows), p'-traces (the same as the p-traces but with losses and delay), s'-traces (the same as s-traces but with losses and delay) and finally v'-traces (with an associated encoding and delivery quality). The interaction between the application domain and 5G System relies on QoS Flows. The document also introduces the framework of Split Rendering with Asynchronous Time Warping (ATW) Correction. In this scenario, an XR Server utilizes an XR engine to formulate an XR Scene based on data from an XR device. XR Server handles XR viewport rasterization and pre-rendering, while the device manages real-time pose correction. By dividing the XR graphics workload between the XR server and device, motion-to-photon latency is minimized through on-device ATW or similar techniques. Edge Computing enables the deployment of cloud computing capabilities in proximity to the cellular network, reducing latency and backhaul traffic. This plays a critical role in ensuring optimal performance in 5G systems for CG and XR [56].

TR [57] introduces Edge Computing as a network framework that facilitates XR and CG. By deploying cloud computing facilities and service surroundings in proximity to the cellular network. Edge Computing offers advantages such as increased bandwidth, reduced latency and minimized backhaul traffic. It serves as significant enabler for 5G systems in succeeding the necessary performance for XR and CG. The TR also discusses the Generic DL traffic model, which consists of two subcategories. The first is the single stream Downlink (DL) traffic model, representing XR DL traffic as a succession of video frames reaching at gNB at various frame rates and with unpredictable jitter. The frame sizes follow a specific distribution. The second is the multi-streams DL traffic model, which includes options like P-frame + I-frame (Group-Of-Picture (GOP) or slice-based traffic pattern), audio/data and video and FOV with omnidirectional stream. For the Generic Uplink (UL) pose/control traffic, packets are periodically transmitted to the UE with specific parameters.

In this TR [57] there are proposed also some scenarios regarding deployment, capacity evaluation, power consumption evaluation, mobility and coverage evaluation. In terms of deployment, different types of applications exhibit varying thresholds for the roundtrip response time. Ultra-Low-Latency applications necessitate a maximum delay of 50 ms, while Low-Latency applications require a maximum delay

of 100ms. Moderate latency applications, on the other hand, have a maximum procrastination requirement of 200ms. Non-critical latency applications can tolerate delays exceeding 200 ms. In the context of XR deployment, three distinct scenarios are considered. The first is Dense Urban, where XR UEs are located in urban areas with densely implemented gNBs, maintaining an inter site distance (ISD) of 200m. The second scenario is Indoor Hotspot, focusing on indoor XR users engaged primarily in VR or CG applications for work and gaming. Lastly, Urban Macro entails a larger ISD at about 500m, accommodating XR users spread across a wider area. Given the extensive ISD in this deployment, XR deployments with lower data rates become more relevant [57].

When evaluating capacity, the satisfaction of a UE is determined by meeting the Packet Error Rate (PER) and Packet Delay Budget (PDB) requisites for all its streams. There are specific directions to consider in this regard. In the evaluation focused solely on the DL, only DL streams are taken into consideration to determine UE satisfaction. Similarly, in the evaluation focused solely on the UL, only UL streams are considered for assessing UE satisfaction. System capacity, a key performance indicator (KPI) for capacity studies, is defined as the peak user capacity per cell with a satisfaction level of at least 90% or 95%. 5G NR can support AR, VR and CG applications, with the highest capacity in dense urban and indoor hotspot scenarios. Higher data rates and stricter PER requirements result in lower system capacity for AR, VR and CG applications, while larger system bandwidths increase capacity [57].

As regards XR UE power consumption evaluation, UE power consumption is one KPI for power estimation and Power Saving Gain(PSG) is calculated by comparing power utilization to the Always On baseline. High PSG can lead to a lower percentage of satisfied UEs, particularly for CDRX. Therefore, power saving gain should be considered along with power consumption. Power saving techniques include CDRX, MIMO layer adaptation, cross slot scheduling and PDCCH supervision adaptation [57]. Increasing application generation and data rate lead to higher UE power consumption. Reducing pose periodicity can help decrease power consumption. The choice of CDRX configuration affects power saving gain and capacity trade-off [57].

Regarding XR coverage evaluation, it depends on various factors such as link direction, bit rate, power, etc. The KPI used to assess coverage is XR signal reach, defined as the coupling gains 5% in the Cumulative Distribution Function (CDF). UL coverage is typically worse than DL coverage in dense urban and urban macro deployment scenarios [57]. Concerning XR Mobility, higher PDB conduces to better mobility KPIs, while higher frame rate leads to worse number of consecutive XR packets lost. Lower handover interruption time than PDB results in lower mobility KPIs, whereas higher handover interruption time conduces to poor mobility KPIs [57].

In this TR [58], it is presented a Client reference model and client architecture for VR QoE measurement. The VR application and Dynamic Adaptive Streaming over HyperText Transfer Protocol (HTTP) DASH access engine work together to provide a smooth streaming experience. The VR application uses sensor information and metadata to decide which adaptation sets to use, while the DASH access engine fetches media segments for those sets and adapts as needed for network bitrate. The DASH access engine has no knowledge of the viewport, while the VR application does not handle media fetching. In this reference model, monitoring locations exist where metric data can be collected for the purpose of Computation function and Metrics Collection. These points may not always directly interface with the Metrics collection and computation (MCC) and the MCC may be included as part of the VR application [58].

The access mechanism retrieves the Media Presentation Description (MPD) and media segments based on the specified order of the VR application. The file decoder processes the VR Track, while the sensor captures the user's pose to generate the viewport. Utilizing decoding signals, rendering metadata, pose and FOV, the VR Renderer determines and renders the suitable viewport. The VR application manages the file decoder, access mechanism and rendering derived from media control and user pose data. The VR QoE client reference model recognizes these interactions through sensors and translates them into pose data, including head pose, gaze direction, skeleton and hand gestures. The latency of VR interactions, which impacts the user experience, relies on packing, encoding delivery technologies and device competencies. Single-stream region-independent streaming delivers VR content as a single stream with consistent quality settings across a 360-degree view, while streaming methods and diversified stream region-dependent encoding utilize pose data to determine the appropriate stream with superior quality region linking to the viewport [57]. The technical report (TR) also presents an Immersive media metrics client reference model consisting of five operational modules: sensors, network access, media playback, client controller and media processing. Metrics computing and reporting (MCR) component calculates various metrics by extracting measurable data from the functional modules [58].

Furthermore, another TR [59] introduces VR audio application scenarios supporting two-way communication with specific constraints, including low mouth-to-ear latency requirements for conversational proficiency. Various aspects such as the audio rendering system, audio production format, audio capture system, audio storage format, audio distribution format and content production workflow are discussed. A comprehensive workflow for content creation and distribution in VR is also presented, incorporating channel-based audio, object-based and scene-based representations.

In these TR and TS [59], [60], it is presented an architecture for VR streaming services. The process for distributing VR content using DASH-based services involves several steps. First, the content is acquired and uploaded for preparation, and metadata is made available for encoding and file format encapsulation. Content is initially manipulated to establish

mappings between media components and the 3GPP three-degrees of freedom (3DOF) system. It is then subjected to encoding, ensuring the proper encoding of 3D audio and spherical video signals. In the case of file-based distribution, the 3GPP VR tracks are combined into a identical file.

Alternatively, for DASH-based delivery, the content is mapped to DASH segments, generating appropriate adaptation sets. Delivery can be accomplished through file-based or DASH-based methods, leveraging 3GPP services such as DASH in Packet-switched Streaming Service (PSS) or DASH over Multimedia Broadcast Multicast Service (MBMS). The content becomes available for delivery after media encoding, as a complete file or segmented tracks optimized for DASH. Multiple versions of equivalent content may be encoded to cater to different viewpoints. At the receiving end, a VR application interacts with various functional blocks within the receiver's VR service infrastructure, including media decoding, delivery client, rendering environment, viewport sensors and file format encapsulation. The entire operation is dynamic, utilizing pose information in rendering units, DASH client or download for delivery and for decoding enhancements. Inverted operations are executed and communication remains dynamic, particularly with changing sensor metadata. Delivery client establishes communication with file format engine, while diverse media receivers decode data and offer input to the rendering process [59], [60].

In this TS [60] they are also described five audio rendered definitions, including the Reference Renderer and the Common Informative Binaural Renderer (CIBR). The Reference Renderer offers an audio rendering workaround for a relevant media profile, supporting binaural and loudspeaker based rendering and diegetic/non-diegetic content. The CIBR is a binaural renderer with four components: Equivalent Spatial Domain (ESD) to Higher Order Ambisonics (HOA) conversion, sound field rotation, HOA to binaural rendering and a mixer for diegetic/non-diegetic content. The External Renderer enables alternatives to the Reference Renderer and supports binaural/loudspeaker based rendering and diegetic/non-diegetic content. The Common Renderer API empowers the utilization of an External Renderer that is capable of supporting all VR-Stream media configurations. It incorporates standardized implementation descriptions specified in 3GPP technical requirements. The External Renderer API enables the adoption of an External Renderer and provides essential data for establishing a seamless connection between a VR Stream media profile and an External Renderer.

In TS [60], Video Media Characteristics are detailed. Basic Video Media Profile enables the downloading of elementary streams for VR data created using Basic Operation Point H.264/AVC. It also facilitates adaptive streaming of VR video material by providing multiple selectable Representations within a single Adaptation Set in a DASH MPD. The Primary Video Media Characteristic enables the downloading and streaming of elementary streams for VR content created using the H.265/HEVC Main Operation Point. It also enables adaptive streaming of VR video content by offering multiple selectable Representations within a unique Adaptation Set in DASH MPD. Advanced Video Media Profile enables for VR content the downloading and streaming of elementary streams created using the Flexible operation point H.265/HEVC. It also permits unrestricted utilization of rectangular region-wise packing, stereoscopic and monoscopic spherical video until 360 degrees.

In this TR [61], a diagram is presented that offers a fundamental synopsis of the key functionalities of an AR device. Primary capabilities of AR/MR technology encompass: a) AR/MR applications that seamlessly integrate audio and visual data into user's surroundings b) Media Access Function, which allows entry to media and other AR content required for AR Runtime or the scene manager c) AR Runtime, which interfaces with a framework to execute essential functions such as retrieving controller and peripheral state, tracking positions, performing spatial computing and rendering frames d) Scene Manager, which assists application in organizing the spatial and logical representation of a multisensory scene with support from AR Runtime e) Peripheral devices, including sensors, cameras and displays, that establish a physical association to environment [61].

XR Spatial computing involves sensor data analysis to generate insights about three-dimensional environment surrounding an AR user. It encompasses 3D reconstruction, spatial mapping, localization and semantic perception. Spatial computing can be executed on the AR or XR device or utilizing cloud/edge resources (Fig. 4). Two scenarios exist: 1) spatial computing with storage and retrieval facilitated by an XR Spatial Description server and 2) offloading compute functions to an XR Spatial Compute server. These scenarios, known as the STAR and Edge-dependent AR (EDGAR) architectures respectively, depend on the device's processing power and the complexity of XR compute functionalities [61]. Latency plays a vital role when rendering occurs at the edge or cloud, demanding low-latency and high-quality data delivery across the network. AR Runtime manages buffers for eye, depth and sound and careful execution of processes is necessary to meet E2E latency requirements. Rendering closer to the AR UE simplifies meeting latency demands [61].

In this TR [61], the Media Access Capability for AR is also introduced. It enables the AR user to access and stream media content. The included functions, essential for both uplink and downlink, comprise the following: a) Media Session Handler: a versatile task responsible for configuring 5G System capacities, which should involve setting up edge operations, offering QoS assistance and facilitating notification b) 5G connectivity: encompassing a modem and 5G System modules enabling the UE to associate to a 5G network and retrieve the modules and services provided by 5G System. c) Content Delivery Protocol: a container format and protocol for delivering media content among network and User Equipment, aligned with application requisites.

d) Content protection and decryption: handling the protection of content to prevent unauthorized playback on unapproved devices. Additionally, there are buffers, typically managed by the AR Runtime, such as eye, depth and sound buffers. The proximity of rendering to the AR UE influences the ease of meeting latency requirements. e) Codecs: employed for compression and decompression of rich media [61].

The AR Framework, established by the ETSI Industry Specification Group, has devised a prototype structure for AR systems, comprising three layers: hardware, software and data. The hardware layer encompasses processing components, rendering interfaces and tracking sensors. The software layer comprises a vision and a 3D rendering engine, and finally, the data layer incorporates interactive content and world knowledge. The framework outlines the attributes of an AR system and delineates its functional components [61].

Lastly, the Moving Picture Experts Group (MPEG) has devised an architectural framework to provide guidance for immersive media and scene depiction. The focus of the MPEG-I architecture centers on utilizing buffers as a medium of data transfer across the rendering pipeline and media access. The architecture incorporates defined interfaces, including the Media Access Function API, which enables the retrieval of media referenced by the scene description through buffers [61]. The approach of 5GMS, which involves the segregation of data and control planes, proves advantageous for the incorporation of AR services into 5G networks. By adopting this approach, third-party services can leverage 5G functionalities such as split rendering and real-time communication. Control plane aligns with the principles of 5G Media Streaming (5GMS), while the media plane remains generic, accommodating diverse operators and third-party services. This expansion of 5GMS principles facilitates the integration of various services into the 5G system [61].

B. 3GPP SOLUTIONS

There was a big effort within 3GPP last years to provide solutions for running XR services over 5G mobile networks. Table 2 enumerates the aforementioned factors based on different sections of the network, such as radio access, core network and other relevant components.

C. LIST OF KPIS

XR services require a new set of KPIs that are presented in Table 3, where they present the downlink and uplink performance requirements for various immersive architectures like MR,AR,VR and CG as part of the evolutionary progression of 3GPP standards in 5G networks. They also outline key parameters and metrics that are essential for ensuring a seamless and responsive user experience during DL and UL data transmission. The tables differentiate between the performance requirements for AR, VR, MR and CG to address their demands and distinct characteristics. They also include metrics such as the minimum data rate, packet loss, latency and reliability that are crucial to deliver seamless and immersive XR experiences from the user's perspective. Video frame size and jitter are generated from Truncated Gaussian Distribution. All immersive services on the uplink include an uplink flow responsible for transmitting from UL pose repeated minor control packets. The maximum allowed delay for this type of flow is relatively small, around 10 milliseconds.

These estimations were derived from system-level simulations in a multi-cell environment. Furthermore, it is universally acknowledged that the power and capacity of UE are jointly evaluated to prevent the implementation of any power enhancements that could lead to decreased performance [57]. These immersive technologies aim to enhance user experiences by merging virtual elements with the real world or creating entirely virtual environments. VR systems generally offer a FOV between 100 to 150 degrees, allowing users to perceive a wide immersive environment. The FOV for AR and MR systems is 20 to 50 degrees and it depends on the specific device or application. Low latency is crucial in all immersive systems to provide a seamless and responsive experience. AR and MR systems strive for latency 15 and 10 milliseconds and as shown in Table 3, some low-interactive VR can tolerate around 1000 ms latency. This is because the HMD can use a buffer to save multiple frames and play them with a certain delay and it can also effectively address network jitter. The buffer size and delay can be determined by the specific application's latency tolerance. VR systems often rely on wired connections to ensure high-speed and stable data transfer between the computer and the headset. AR and MR systems can operate using either wired or wireless connections, depending on the specific device and application. While VR systems do not require specific data transfer rates (uplink 150-200 kbps), AR and MR systems typically aim for higher rates between 0.02-1 gigabits per second (Gbps) to ensure smooth transmission of augmented or mixed reality content.

SA4 (Service and System Aspects) is now still working on the traffic characteristics of XR type services, the results of which should be used as the baseline for traffic models design in RAN1 evaluation. In the Table 4, some initial considerations for the XR traffic model design are given. Split rendering, Viewport independent streaming and CG deployments necessitate a maximum data transfer rate of 100 megabits per second. If the XR service uses a higher resolution and frame rate higher traffic throughput is required. PDB in Table 4 indicates the maximum allowable time delay between the Policy Charging Enforcement Function (PCEF) and the wireless device, which restraints the maximum packet transfer delay. Round-Trip Time (RTT) affects also the QoE. As XR services become more dynamic and complex the latency requirements increase. Ensuring minimal latency is crucial to ensure a seamless user experience without any noticeable delays. PER serves as a measure of reliability, representing the percentage of packet faults occurring from E2E at the application layer within the PDB

TABLE 2. List of 3GPP solutions for XR services in 5G mobile networks.

Category	Solution
Core Network	• Policy control improvements to support multi-modality flows coordinated 5G transmission [62]. • Achieving application synchronization and QoS policy coordination for Multi-modal Data flows among multiple UEs [62]. • Mechanisms that enable code/rate adaptation to meet requirements for services. 5GS Information exposure for XR/media enhancements [62]. • Enhancement of QoS model and policy control for PDU Set integrated packet handling in both the downlink and uplink direction (in UE, RAN and/or UPF). Which types of PDU Set shall be supported for PDU Set integrated packet handling by 5G network [62]. • Support of varied QoS handling estimating different importance of PDU Sets by treating packets (PDUs) belonging to less important PDU set(s) and enhancement of the QoS model and policy control for the importance/dependency information associated with a given PDU set [62]. • Efficient coordination of UL and DL transmissions is implemented to fulfill Round-Trip latency requirements. This involves possible interactions between the AF and the 5th Generation System (5GS), along with potential QoS enhancements [62]. • Improvements in policies are implemented to minimize jitter in QoS flows that support XR and media services. These enhancements primarily concentrate on requirements provisioning from the AF and the extension of Policy and Charging Control (PCC) rules [62]. • Enhancements to power savings for XR services (how to improve power management schemes like CDRX (cycle, on duration and inactivity timer) to achieve the best trade-off among KPIs like latency and device battery lifetime [62]. • Support of Trade-off of QoE (throughput/latency/reliability) and Power Saving Requirements (device battery life) [62].
Application	• A Spatial Computing Server is employed in conjunction with XR-related devices to oversee and retain data pertaining to the physical environment and the positioning of XR entities [54]. • The power consumption could be decreased by reducing pose periodicity [56], [57]. • Raster-based split rendering (to distribute the rendering workload across multiple GPUs, servers, or nodes) is typically used to achieve high performance and scalability in large-scale XR systems [56]. • DASH over MBMS extends the capability to broadcast networks by allowing the DASH client to dynamically adjust to fluctuating network statements in real-time. This results in a more efficient and robust delivery of multimedia content over the broadcast network, providing a better user experience for viewers [59], [60].
RAN	• The incorporation of novel functionalities and improvements (such as downlink interruption, mini-slot transmissions, grant-free transmissions and front end loaded Demodulation Reference Signals (DMRS)) is motivated by the objectives of Ultra-Reliable Low Latency Communications (URLLC) and power optimization [63]. • The disparity between the periodicity of PDCCH monitoring and XR traffic periodicity is addressed to minimize latency and ensure that Ultra Low (UL) packet transmissions align with the PDB requirements [63]. • Enhancements like support of new Downlink Control Information (DCI) for higher reliability, lower latency and resolution of traffic conflicts [63]. • Network slicing is an additional valuable method for effectively handling a wide range of diverse XR services [54]. • 5G NR could be supported, where capacity in urban macro scenario is inferior than that in dense urban and indoor hotspot scenarios, in deployments with uplink video. NR system capacity is smaller for applications that require higher data rate and higher with larger PDB value and also the capacity is higher with larger system bandwidth [57]. • Higher PDB conduces to lower and increased mobility KPIs and higher frame rate conduces to higher and smaller number of consecutive XR packets lost. As the handover interruption time increases, the network performance in terms of mobility decreases [57]. • XR-Awareness in both UL and DL conduces to improvements of gNB radio resource scheduling based on the notions of PDU set and Data Burst. There are also proposed some alternatives between the mapping of PDU sets and QoS flows in the NAS and also regarding the allocation of QoS flows to Data Radio Bearers [57].
Edge	• Freshly introduced application layer interfaces specifically designed for Edge Computing that are defined in 3GPP enable the seamless integration of Edge Applications with the Edge AS and ensure their smooth operation and management [54]. • Introduction of edge computing enhancements ((by minimizing the distance between the network-side AS and the application client on the UE side) [54]. • Edge Computing harnesses the capabilities and service environments of cloud computing. When deployed in close proximity to the cellular network, it provides benefits such as increased bandwidth, decreased latency, minimized backhaul traffic and improved overall performance [57]. • When rendering occurs at the edge or cloud, it is essential to transfer the rendered data efficiently over the network, ensuring low-latency and high-quality delivery. The proximity of rendering to the AR UE plays a crucial role in meeting the latency requirements, with closer rendering locations making it easier to achieve the desired latency levels [61].

D. XR SERVICES AND 5G-A NETWORK ARCHITECTURES

Furthermore, 3GPP introduced a set of new XR services architectures that we list and describe below(Table 5). Those XR architectures define the role of XR in the 5G-A ecosystem as we describe in more detail below(Table 5).

In (Fig. 3) it is depicted the application layer of an XR model of 5G as it is defined in 3GPP Rel.18, where different colors indicate the respective element of a specific 3GPP XR service architecture. Encoding and decoding tasks are usually parts of the application layer which handles tasks such as video and audio processing, compression, decompression and rendering. Generally, in XR applications, the client device (XR Device) has the responsibility of presenting virtual content to the user, which means to interpret and convert compressed data, like videos or 3D models, into a format that enables real-time visualization and user engagement.

On the server side (XR Server), there are more heavy tasks. This includes the encoding, which involves the compression of data (videos or audio), so as to optimize the transmission to the client device. Also, the server has some additional responsibilities like organizing content, optimizing network performance and processing data in real-time. The XR device (in UE) handles the processing of tracking data and sensor information, while the complete XR environment is transmitted and decoded for the user's experience (Viewport-Independent delivery, Viewport-Dependent Streaming). In Viewport-dependent streaming, pose information (from pose generator) is used for rendering (Viewport Rendering) and tracking purposes on the

TABLE 3. VR, AR, MR and CG System Specifications and Technical Requirements [11], [40] [56], [63] [64], [65].

Application	CG	VR	AR	MR
Traffic Model (DL)	Video single-stream	Video single-stream	Video single-stream	N/A
Bitrate (DL)	(8 Mbit/s, 30 Mbit/s, 45 Mbit/s) (DL)	(30 Mbit/s, 45 Mbit/s, 60 Mbit/s) (DL)	(30 Mbit/s, 45 Mbit/s, 60 Mbit/s) (DL)	N/A
Packet Rate (DL)	60 fps ([30, 90, 120] fps) (DL)	60 fps ([30, 90, 120] fps) (DL)	60 fps ([30, 90, 120] fps) (DL)	N/A
Packet Delay Budget (DL)	15 ms (DL)	10 ms (DL)	10 ms (DL)	N/A
Packet Error Rate (DL)	1% (DL)	1% (DL)	1% (DL)	N/A
Number of Streams (DL)	1% (DL)	1% (DL)	1% (DL)	N/A
Jitter (DL)	[-4, 4] ms or [-5, 5] ms (DL)	[-4, 4] ms or [-5, 5] ms (DL)	[-4, 4] ms or [-5, 5] ms (DL)	N/A
Periodicity (UL)(except audio and data that is 100fps in both DL and UL)	250fps(pose or controller)	250fps(pose or controller)	60 fps(pose or controller and scene+video)	N/A
Success % (UL)	99 (90, 95 optional)	99 (90, 95 optional)	99	N/A
Delay Bound (UL)	(10, 15, 60 optional) 30	(10, 15, 60 optional) 30	(10, 15, 60 optional) 30	N/A
Data Rate (UL)	0,02 Gbps	150kpbs-200kpbs	0.02 - 1.0 Gbps	0.02 - 1.0 Gbps
Jitter (UL)	No jitter	No jitter	[-4, 4] ms or [-5, 5] ms (DL)	N/A
Screen	N/A	Occlusion	Translucent	Translucent
Display	N/A	HMD(Head-Mounted Display))	OHMD (Optical Head-Mounted Display)	OHMD (Optical Head-Mounted Display)
Environment	N/A	virtual	Passive virtual & real	Passive virtual, active virtual & real
Data Rate (DL)	N/A	0.02 - 1.0 Gbps	0.02 - 1.0 Gbps	0.02 - 1.0 Gbps
Latency	N/A	20-1000 ms	20 ms	10 ms
Refresh Rate	N/A	$\sim$ 90 Hz	$\sim$ 90 Hz	$\sim$ 90 Hz
Pixels-per-Degree	N/A	10-15	30-60	30-60
Field-of-View	N/A	100○ - 150○	20○ - 50○	20○ - 50○

UE to ensure proper alignment and synchronization of virtual content [54].

Additionally, XR rendering metadata (Generalized XR Split Rendering) provides additional information about the rendering process and characteristics of the XR content, which can be used to optimize the rendering specifically for the viewport.Viewport pre-rendering is a technique where a scene or frames are rendered in advance, to reduce the computational load on the target device while the device does pose adjustments asynchronously to accommodate pose changes (Raster-based split rendering-ATW and Generalized XR Split Rendering). Routing and network media processing between XR devices and servers involve the transmission and manipulation of media data across a network connection and play a crucial role in ensuring seamless and reliable communication between XR devices and servers (XR conversational Architecture, XR Conferencing Architecture) [54]. The scene generator is responsible for creating and rendering the virtual environment within the XR runtime. It uses various inputs, among them the pose information that is provided by the pose generator, to accurately position and align the virtual objects and scenes in relation to the user's movements and orientation. With pose information, the XR runtime can reassure that the virtual content is properly synchronized with the user's position and movements, creating a realistic and immersive XR experience (XR Distributed Computing Architecture). The pose generator always has a connection with controllers in context of XR experiences. Sensor Data Processing in 5G XR involves analyzing data captured by sensors to enhance the XR experience. It enables real-time tracking, gesture recognition and spatial awareness (XR Distributed Computing Architecture). Integration with 5G capabilities like low latency and edge computing further enhances performance and user experience. Controllers (handheld devices or motion controllers) are used to interact with objects in an XR environment.

Media decoders within the XR device are responsible for decoding the media content, while the rendering of the viewport occurs directly without relying on specific viewport information. Content protection and decryption are functions that are primarily responsible for safeguarding digital content from unauthorized access and playback on unauthorized

TABLE 4. Traffic Attributes for various XR architectures [54], [66] [67].

Architecture	DL Rate range	UL Rate range	DL PDB	DL PDB	RTT	DL PER range	UL PER range	Traffic periodicity range	Traffic file size distribution
Viewport independent streaming	100 MBPs	HTTP requests in each second. TCP handshake	200ms-300ms	200ms-300ms	N/A	10e-6	10e-6	Nearly stable	Nearly Stable
Viewport dependent streaming	25 MBPs	More frequent HTTP requests every 100ms. TCP handshake	300ms	300ms	N/A	10e-6	10e-6	Nearly stable	Nearly Stable
Viewport Rendering in Network	from 100MBit/s to 10 Gbit/s	For further study	For further study	For further study	For further study	For further study	For further study	For further study	For further study
Raster-based Split Rendering with Pose Correction	100 Mbit/s	500 kbit/s	20ms	10ms	50ms	10e-4	10e-4	Nearly stable	For further study
Generalized Split Rendering	For further study	For further study	For further study	For further study	For further study	For further study	For further study	For further study	For further study
XR Distributed Computing	For further study	For further study	For further study	For further study	For further study	For further study	For further study	For further study	For further study
XR Conversational	50Mbps	N/A	100ms- 150 ms	N/A	N/A	10e-2- 10e-3	10e-2- 10e-3	For further study	For further study
XR Conferencing	3Mbit/s up to 50Mbit/s per user	3Mbit/s up to 50Mbit/s	Permitting instantaneous orlive communication.	Permitting instantaneous orlive communication.	Permitting instantaneous orlive communication.	For further study	For further study	Nearly stable(with peek during startup)	>50Mb at the beginning
CloudGaming	100 Mbps	1Mbps	2.5ms	2.5ms	5ms	10e-4	10e-4	Nearly stable	Nearly Stable

devices. Content Delivery Model Receiver is typically part of the XR device or application and is responsible for receiving the XR content, which can include various media types such as video, audio, or 3D models, from the network. Its role is to receive, buffer and process the media content [59], [60]. Content decoding model has a direct relation with Content Delivery Model Receiver. It encompasses various techniques, such as video codecs (e.g., H.264, HEVC) or audio codecs (e.g., AAC, MP3), to decompress and decode the media content. File format processing involves handling and interpreting the structure and specifications of a specific file format. XR content can be delivered either through a file-based approach or using DASH-based delivery methods. In file-based delivery, the encoded content is provided as a complete file.

Alternatively, for DASH delivery, the content is divided into smaller segmented tracks. This delivery can be facilitated through 3GPP services like DASH in PSS or DASH-over-Multimedia Broadcast Multicast Service (MBMS) [59], [60]. The DASH Access Client uses the DASH MPD file to request and receive the content in a segmented manner from the DASH Delivery system. The DASH Integration ensures the seamless integration of these components, allowing for adaptive streaming and efficient delivery of multimedia content over HTTP. Multiplexing and demultiplexing enable efficient transmission and extraction of multiple data streams within the file. File format encapsulation on the server prepares the XR content for delivery by packaging it into a suitable file format. It is proposed to use spatial compute functions in UE to enable real-time spatial processing and interactions [61]. The UE is responsible for executing spatial compute algorithms, such as object tracking, spatial mapping and gesture recognition, in real-time to enhance the XR experience. By performing these computations locally, the UE reduces the reliance on external resources and enables faster response times. A spatial description server functions as a storage and retrieval system for spatial information and spatial data in the context of XR experiences [61].

The XR device can query the spatial description server to retrieve relevant spatial information, such as 3D models or

TABLE 5. Core functionalities per XR architecture [54], [61] [68], [69].

XR Architecture	Functionality
Viewport-Independent Delivery	• Content is delivered without considering the specific viewport or display device. • Same content is eligible to be delivered to different devices with varying capabilities. • Content can be pre-rendered or rendered in accordance with the competencies of the client device and in real-time.
Viewport-Dependent Streaming	• Content is streamed and rendered based on the specific viewport or display device. • Adaptive streaming techniques are utilized to enforce the quality of the content by adjusting it according to the network conditions and capabilities of the UE. • Optimizes efficient usage of bandwidth while providing a better experience by customizing the content to fit the specific viewport.
Viewport Rendering in Network	• Rendering is partially or completely handled on the network or server-side. • Reduces the processing demands on client device. • Permits the use of less powerful devices or thin clients to access XR content.
Raster-Based Split Rendering	• Rendering process is divided between client device and server. • Client device renders the foreground, while the server renders the background. • Reduces the computational load on the client device while maintaining visual quality.
Generalized XR Split Rendering	• Similar to Raster-Based but applies to a broader range of scenarios. • Allows for more flexible distribution of rendering tasks between client and server.
XR Distributed Computing	• Leverages the power of distributed computing resources to manage demanding XR processing tasks effectively. • Distributes the computational workload across multiple devices or servers, resulting in enhanced performance. • Enables physics simulation, real-time rendering, or other computationally intensive operations.
XR Conversational Architecture	• Enables XR experiences that leverage natural language processing and voice-based interactions, enhancing user interactions within the XR environment. • Incorporates conversational agents or virtual assistants seamlessly into the XR environment. • Offer the possibility to users to interact with the XR content through voice commands or natural language inputs.
XR Conferencing Architecture	• Designed for collaborative XR experiences and virtual meetings. • Facilitates the participation of multiple users in a shared XR environment, allowing them to collaborate and engage with shared content together. • Offers capabilities such as avatars, spatial audio and synchronized interactions, enhancing the immersive and interactive experience for participants in the shared XR environment.
Cloud Gaming	• Enables end-users to interact with video games remotely without necessitating dedicated gaming hardware. • Offers compatibility with a diverse array of devices, encompassing smartphones, tablets, smart TVs and computers. • Offers instant access to a library of games without the need for downloads or installations.

reference points, to augment its perception of the environment and provide accurate spatial context for the user. Metadata generation in 5G XR on Server involves creating descriptive data that enhances XR content, assists in content adaptation, enables content management and provides valuable analytics for improving user experiences and optimizing resources (XR Distributed Computing Architecture). XR Engine in the server focuses on processing XR content and generating frames, while the XR Runtime on the client device interfaces with the platform and handles input-related operations. It provides interfaces or APIs that can be accessed by a 5G-XR Aware Application. The latter consists of two subroutines, XR Session Handler and XR Engine [54].

The XR Session Handler interacts with the 5G-XR AF (network layer) to initiate and manage the provisioning of an XR session. Additionally, it provides APIs that enable the 5G-XR Aware Application to utilize its functionalities [54]. XR Session Handler sometimes sets up edge operations, provides QoS support and support monitoring. On the other hand, the XR Engine communicates with the 5G-XR AS (network layer) to access XR-related data and capacities like sensor integration, tracking, data analysis and communication

with the XR Session Handler to oversee XR session management. By controlling the 5G-XR Engine, the 5G XR Aware Application can tailor the XR experience according to its specific requirements, providing a more customized and optimized XR session for the end-user. The content acquisition and preprocessing stages prepare the content, which is then encoded using a specific content encoding model. The encoded content is transmitted by the content delivery model sender, while the content model provides the necessary instructions and information for rendering and displaying the XR experience. Capture process involves capturing raw data or input from multiple sources, such as cameras or sensors. Sphere stitching is a technique that is used to merge and stitch together the captured data. Once the data is stitched, projection techniques are applied to transform the spherical representation into a format suitable for rendering on a flat display or viewport. Packing refers to the process of organizing and packaging the transformed and projected data into a suitable file format or container for efficient storage [59]. The interface between the application domain and the foundation of the 5G System relies on QoS Flows.

Data network (part of 3GPP model of Media Streaming in 5G) is the network infrastructure that facilitates the transmission of data between different components and entities within the 5G XR system. Within this scheme (Fig. 4), the application provider, 5G-XR AS and 5G-XR AF may be represented as entities or components involved in the delivery and management of XR services. The DN would indicate the network connectivity and pathways through which these entities communicate and exchange data within the overall architecture. 5G-XR Application Provider is an external entity responsible for delivering XR applications and services by utilizing the capabilities of 5G-XR clients and networks. Its role is to offer an immersive XR undergo to the 5G-XR Aware applications, leveraging the features and functionalities provided by the 5G-XR ecosystem. A 5G-XR AS plays the role of an AS that accommodates 5G-XR media content and associated media functionalities [54].

The 5G-XR AF assumes the role of providing a range of control capabilities to the XR Session Handler on both the 5G-XR Application Provider and the UE [54]. It is capable of initiating requests for customized treatment from the PCF or engaging with other NFs such as Network Exposure Function (NEF). PCF generates the QoS monitoring policies for the data rate measurement, the normal data transmission interruption event measurement and the congestion information measurement [62]. In the context of 5G-XR, the NEF can interact with the 5G-XR AF to facilitate the exchange of information and enable the AF to access network resources and functions as required by XR services. The 5G-XR AF plays a vital role in enabling efficient communication and coordination between the various components of the network to ensure optimal performance and functionality for XR sessions. The Network Data Analytics Function (NWDAF) in XR architecture collects and analyzes data on XR sessions, user experiences and network performance to provide insights into XR-specific patterns like application usage, latency, bandwidth and QoE. These insights optimize XR service delivery, enhance application performance and improve the overall XR experience [62]. In general, NWDAF predicts the estimated QoS according to the collected information and performs the judgement of whether the trigger conditions are met.

The combination of ASs and edge processing provides significant advantages for 5G-XR applications. As it is referred in [54] in devices that have limited power and resources, edge computing can be used to assist or divide the workload across the network. In this process, the UE acquires sensor data and transmits it to the cloud-side promptly. Upon receiving the sensor data, the cloud-side undertakes essential computational tasks, such as rendering, to produce multimedia data. Subsequently, the processed data is transmitted back to client devices for presentation. This method facilitates the offloading of computationally demanding tasks to the cloud-side, ensuring prompt delivery of processed data to the user's devices. Additionally, it provides advantages such as reduced latency, backhaul traffic and enhanced bandwidth [57].

3GPP has defined application layer interfaces specifically for Edge Computing, enabling the discovery, exposure and administration of Edge Applications, consequently by enhancing the performance and capabilities in 5G-XR applications. Edge XR spatial compute functions refer to processing capabilities executed at the network's edge to support spatial processing and analysis in XR applications [54], [57], [61].The below scheme (Fig. 4) clearly depicts the presence of an Edge XR Spatial Compute Server. These functions involve processing spatial data, such as location, orientation and environmental information, in real-time to enhance XR experiences. By offloading spatial compute tasks to the edge, latency is reduced and XR applications can leverage the localized processing power for faster and more efficient spatial computations. XR spatial description edge cache refers to the practice of storing spatial data at the network edge to enhance the performance of XR applications. Spatial data includes details about the physical environment, such as 3D models, textures, maps and other related elements.

By caching this data at the edge, XR applications can access the required spatial information rapidly and effectively, eliminating the need for repeated data transmission over the network. This edge caching approach minimizes latency, accelerates the loading speed of XR content and delivers a seamless and immersive XR experience to users. In XR Spatial Computing, processing can be done solely on the AR device or augmented with cloud or edge resources. There are two scenarios to consider: a) AR devices carry out spatial computing tasks, while an XR Spatial Description server manages the storage and recovery in XR Spatial Descriptions. b) Certain spatial computing functions are delegated to an XR Spatial Compute server, especially in cases where the device possesses limited processing capabilities or when intricate XR compute tasks are involved. These scenarios are referred to as the STAR architecture

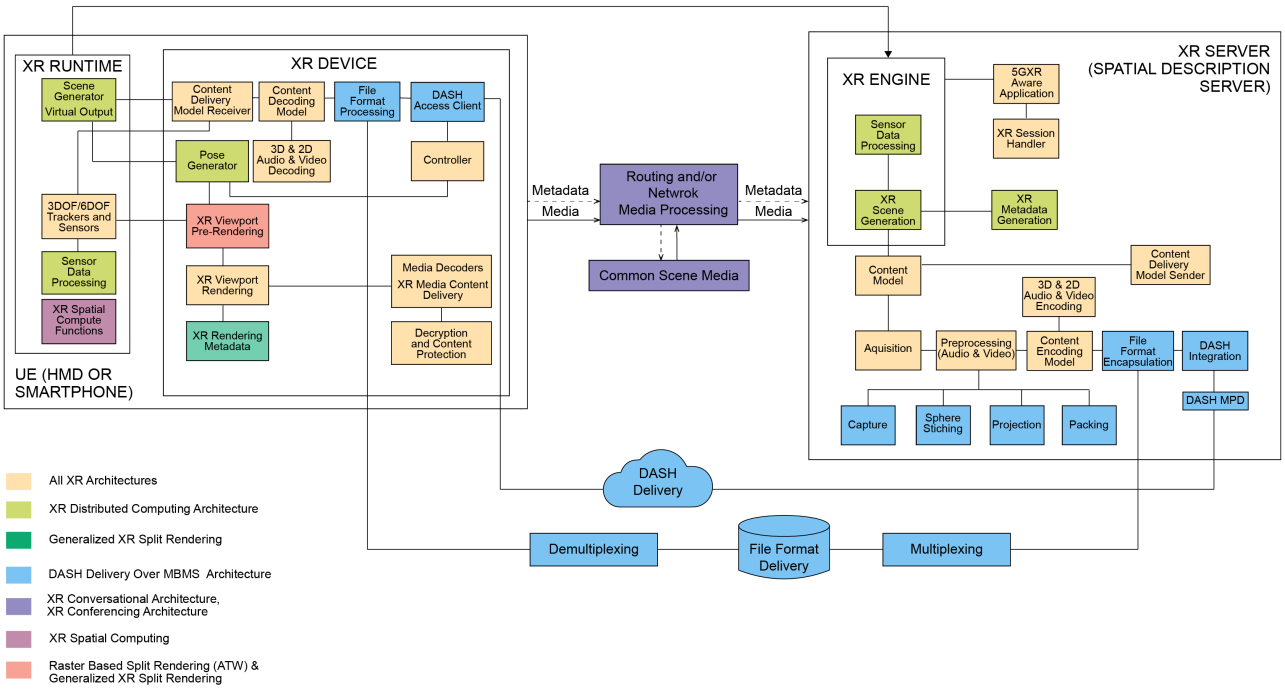

FIGURE 3. 3GPP Rel.18 XR services architectures.

(AR device-based computing) and the EDGAR architecture (edge-assisted computing) respectively. The distinction lies in the degree of reliance on local processing versus external resources [61]. When Cloud edge and XR converge in the 5G ecosystem, it enables powerful capabilities. Cloud edge computing brings computing capabilities and storage in close proximity to XR devices, reducing latency and enabling real-time processing of XR content. This allows for complex XR applications with high-quality graphics and interactive experiences [54], [57], [61].

The 5G QoS model is an integral feature of network architecture in 5G, primarily situated within the core network. It plays a vital role in managing and regulating data flow, employing QoS mechanisms to guarantee diverse service quality levels for applications and user traffic, which is the reason that is described for immersive services and XR. This model establishes guidelines and mechanisms for categorizing, prioritizing and managing various types of network traffic based on their specific QoS demands. By adhering to these principles, the network ensures efficient handling of data and optimized service delivery for different applications.

Every QoS flow is assigned a QFI that uniquely identifies it within the network. User plane traffic that is assigned to a particular QoS Flow within a PDU Session is consistently managed and forwarded in accordance with the corresponding QoS parameters, ensuring uniform treatment throughout the flow [54]. 5G QoS model is crucial for XR architecture and immersive services. Its key roles include: Network Performance and parameters like data rate, latency, reliability and availability to ensure a high-quality XR experience [70] and Dynamic Resource Allocation which enables the allocation of bandwidth and prioritization of XR traffic based on application requirements. Although this TS [71] does not specifically mention XR, it provides insights into the QoS mechanisms and resource allocation considerations for delivering multicast content, which could be relevant for XR applications like E2E QoS Management which ensures consistent QoS across the network path, from XR device to server or cloud [70] or Service Differentiation that supports classifying XR applications into service classes for prioritized treatment, even in congested networks. In TS [70], it is outlined the framework for defining QoS classes, policies and parameters that enable service providers to differentiate their offerings based on QoS requirements and also the interfaces and protocols involved in QoS control and management within the 5G system. And finally, scalability and flexibility that adapt QoS parameters based on network conditions, user needs and application demands [70].

A QoS Flow can be prearranged or implemented using the PDU Session Establishment process. It is defined by a QoS profile delivered by the Session Management Function (SMF) to the RAN, one or more QoS rules and UL and DL Packet Detection Rules (PDRs) delivered by SMF to User Plane Function (UPF). The default QoS Flow, linked to the default QoS rule, is necessary for a PDU Session and remains active for its entire duration. Other QoS Flows can be categorized as Guaranteed Flow Bit Rate (GBR) or Non-GBR, relying on their respective QoS profiles. The QoS profile, containing QoS parameters, is transmitted to

the RAN. Additional QoS Profiles can be supplied for GBR QoS Flows with Notification control activated. The UE classifies and labels UL user plane traffic based on QoS rules, that could be clearly offered, pre-configured, or obtained through Reflective QoS. The SMF associates PCC rules with QoS Flows delivered from QoS and service requisites. For each new QoS Flow, the SMF allocates a QFI and derives its QoS rules, QoS profile and UPF instructions from the associated PCC rules and relevant information offered by the PCF [70].

In the downlink direction of the network, the UPF takes charge of organizing incoming data packets. It applies a marking technique known as N3 User Plane marking, which assigns a classification to the User Plane traffic correlated with a specific QoS Flow. This classification is conveyed through a QFI. Conversely, in uplink direction, when the UE transmits data packets to the network, the UE itself evaluates the packets. Specifically, for an IP-based PDU Session, the UE compares the uplink packets against the UL Packet Filters contained within the Packet Filter Set defined in the QoS rules. This evaluation assists in determining the suitable treatment and handling of the uplink traffic that is based on specified QoS requisites [64].

Enhancements in gNB radio resource scheduling aim to improve the distribution of network assets that have as a base the concepts of PDU sets and Data Bursts [70]. These improvements focus on both uplink and downlink awareness, leading to more efficient utilization of radio resources. To reduce errors in Buffer Status Reports and improve the knowledge of buffered data delay, new Buffer Status tables are proposed. These tables provide more accurate information about the data waiting to be transmitted and for enabling better resource allocation decisions. Regarding the mapping of PDU sets and QoS flows in the NAS and the transformation of QoS flows into Data Radio Bearers, alternative approaches are being considered [71]. These methodologies aim to optimize the allocation of resources tailored to the distinct demands and attributes of various QoS flows. In the DL, the UPF assigns data packets to specific QoS flows relied on the classification rules provided by the SMF. Similarly, in the DL, the gNodeB (gNB) maps QoS flows to Data Radio Bearers. In the UL, the UE maps packets to QoS flows and Data Radio Bearers using QoS rules provided by the Access and Mobility Management Function (AMF) or through reflective QoS techniques [70]. Within 5G XR networks, the NG-U tunnel serves as a specialized tunnel for carrying user plane traffic among the UPF and the gNodeB (gNB) in the 5G XR RAN. The NG-U tunnel is instrumental in delivering improved performance, low latency and efficient handling of user plane traffic in 5G XR deployments. It plays a vital role in ensuring seamless communication and data transfer among UE and core network within 5G infrastructures [70].

The gNB and AMF work together to provide a robust and efficient infrastructure for delivering 5G XR services. The gNB enables wireless connectivity and radio resource management, while the AMF handles the management and control aspects of XR sessions, ensuring smooth and uninterrupted XR experiences for users. AMF is a vital component of the 5G core network, responsible for managing access and mobility aspects for User Equipments (UEs). Its role encompasses crucial modules including user authentication, overseeing mobility management, managing sessions and ensuring QoS enforcement. In the field of 5G XR, the AMF ensures seamless mobility and efficient management of XR sessions as UEs move across different network areas or access points [70]. The SMF oversees session management, including setup, authentication, mobility management and QoS enforcement, for XR devices (UEs) and the network. SMF generates the QoS Monitoring configuration for UPF and RAN: data rate measurement indication, data rate measure frequency and date rate report threshold [62]. Lastly, the PDRs are linked to the UPF in the 5G network. The UPF assumes a critical function in packet routing and forwarding and carries the responsibility of overseeing and controlling the packet drop rates in the network. PDRs are components of the PCC architecture in 3GPP networks. They define conditions for detecting packets within user plane traffic, enabling QoS enforcement, charging and billing, service differentiation and resource allocation. PDRs ensure packets receive appropriate QoS treatment, enforce charging rules, differentiate services and optimize resource utilization. They play a crucial role in implementing policies, identifying packet attributes and providing efficient resource allocation in the network. Overall, PDRs are essential for managing packet-level control and optimizing network performance in 3GPP environments [72]. These informations for QoS flows are described in TSs and TRs related with 5G and are not specifically covered for XR architectures although they are directly related to the XR reality services.

To incorporate XR applications within 5G it is proposed an approach that comply with 3GPP 5GMS framework. The below scheme (Fig. 4) is based on the 5G core network architecture, standardized by 3GPP, which adopts a service-based approach with ‘‘Services’’ as the fundamental building blocks representing the NFs. The Authentication Server Function (AUSF) plays a crucial role in verifying user identities and overseeing the system's mobility features through the Extensible Authentication Protocol (EAP) [70]. It manages the communication among the UE and the AMF for mobility management, while session-related tasks are handled by the SMF [70]. For the UE, the Network Slice Selection Function (NSSF) controls the most suitable network slice, while the Policy Control Function (PCF) establishes rules to govern control plane functions. Functioning as a gateway between the RAN nodes and the DN, the UPF is supervised by the SMF. Storing subscription and authentication data, the Unified Data Management (UDM) supports user management and the AF enables dynamic policy control and charging for applications. Various components involved in user plane communication, including the RAN, UE and NFs, employ dedicated reference points. The service framework encompasses service registration,

discovery and authentication, facilitated by the Network Repository Function (NRF), allowing new services to register and be discovered as necessary [70]. In the 5G service-based architecture, the NWDAF assumes a critical role as a core network function, integrating data-driven AI/ML analytical technologies to enhance intelligent and autonomous network operations and service administration [73]. As it described before NWDAF collects data from various sources, including NFs, AFs and centralized data repository that consolidates data, along with an operation, administration and management (OAM) system, which is utilized for network analytics purposes. It generates statistical and prediction information that contributes to network optimization, performance enhancement and decision-making within the 5G network. The yellow boxes in the Fig. 4 highlight the new elements in the XR enabled 3GPP network architecture.

IV. LESSONS LEARNED AND OPEN CHALLENGES

A. LESSONS LEARNED

Our survey above of a beyond 5G network to run XR services results in several lessons learned from research and practical experience. A set of lessons learned are presented below as we identified from our studies above.

Regarding the power saving techniques, there are some enhancements for XR services. More specifically, improved power management schemes like CDRX (cycle, on duration and inactivity timer) in order to achieve the best trade-off among KPIs like latency and device battery lifetime. C-DRX saves power by periodically turning off UE components. Different sleep modes such as micro sleep, light sleep and deep sleep offer varying degrees of energy savings by selectively turning off specific components. In the context of XR, the traffic pattern follows a quasi-periodic nature with irregular periods and fluctuations around the average arrival time. This irregularity adds uncertainty to the operational duration of the C-DRX technique, which aims to conserve power. Existing approaches heavily depend on frequent Wake-up Signals (WUS) to maintain micro sleep as the only viable sleep mode due to XR's strict requirements on latency. However, a method introduced in reference [74] introduces WUS skipping, allowing XR devices to utilize alternative sleep modes and achieve improved energy efficiency. The results demonstrate a significant energy-saving advantage compared to existing proposals (up to 25% for a 30-fps frame rate).

Moreover, the need to support a trade-off of QoE (throughput/latency/reliability) and Power Saving Requirements (device battery life). Generally, the power consumption could be decreased by reducing pose periodicity and with the introduction of new features (mini-slot transmissions, monitoring adaptation, cross slot scheduling, MIMO layer adaptation, grant-free transmissions, downlink interruption and front-end loaded DMRS for URLLC and power saving reasons). Regarding XR coverage evaluation, the main factors are link direction, bit rate and power. And as concerns XR Mobility, an increased Packet Delivery Rate positively impacts the performance of KPIs, while a higher frame rate negatively affects the number of consecutive XR packets that are lost. An Adaptive ADRX scheme is proposed to enable dynamic adjustment of the duration of active Connected mode DRX and a stochastic optimization problem is constructed to diminish power consumption while adhering to QoS requirements.

Concerning bandwidth, 3GPP with 5G standard has managed to increase data rates and capacity of cellular networks with higher frequency bands, wider channel bandwidths and improved modulation schemes. Due to high demand for large amounts of data in real-time XR applications, cellular networks have enhanced their bandwidth, especially in the transition from 4G to 5G networks with the utilization of high-frequency mmWave spectrum and lower-frequency sub-6GHz spectrum [75]. To comply with these upgrades, network operators should have to upgrade their backhaul and transport infrastructure.

To provide a seamless and responsive experience for XR applications, low latency is essential. Latency could be affected by network congestion, distance and processing delays. 5G networks are designed to provide ultra-low latency, with some cases targeting 1 ms or less and with 3GPP 5G NR standard features like shorter transmission time intervals, faster channel coding and improved handover procedures [76]. Furthermore, inconsistency between the periodicity of XR traffic and the periodicity of PDCCH monitoring and allowance of UL packet transmission to meet the PDB and finally enforcements like support of the new DCI are some methods for higher reliability, lower latency and resolution of traffic conflicts. It is also proposed an SDN architecture to reduce network latency and a MCR scheme for rapid wireless transmissions from numerous EDCs to the UE, minimizing energy consumption. SDN plays a critical role in improving caching and computing in mobile networks, enabling dynamic programming of heterogeneous devices.

There are also policy control improvements to support multi-modality flows in coordinated 5G transmission and improvements of QoS model and policy control for PDU Set integrated packet handling in both the downlink and uplink direction (in UE, RAN and/or UPF). Policy enhancements aim to minimize jitter for QoS flows in supporting XR and media services (targeting on provisioning from AF, which is an extension of PCC rule). Network slicing has been developed to offer to network the probability of being segmented in virtual networks that are enhanced for explicit XR applications. This additional feature provides better bandwidth, latency and security. 3GPP has developed the Service-Based Architecture (SBA) framework, which enables the implementation of network slices for XR services. This framework provides a flexible and scalable approach to network slicing that can be customized so as to meet the specific requirements of XR applications [77].

Another crucial characteristic for XR applications is the concept of edge computing that has been created by Open

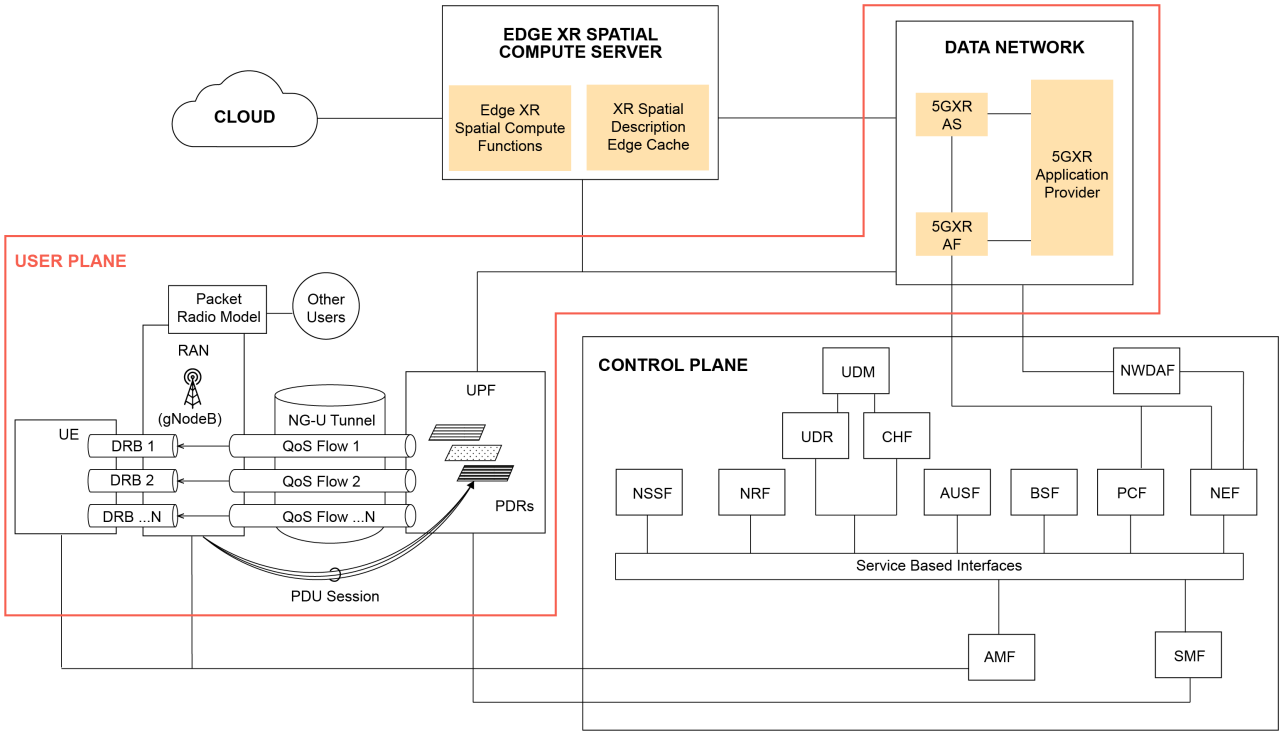

FIGURE 4. XR enabled 5G Advanced network architecture.

Networking Foundation (ONF). This framework architecture is called MEC, as well as in the field of 5G-A era, there are efforts to enhance edge computing so as to bring the AS closer to the UE to achieve real-time responsiveness. Edge computing is crucial for enabling XR and CG and can help with processing in power and resource-constrained devices and ensure their smooth operation and management [78]. Edge Computing empowers the service environments and capabilities of cloud computing. When deployed in proximity to the cellular network, it provides decreased backhaul traffic, enhanced bandwidth and improved performance. As the rendering takes place in closer proximity to the AR UE, the more readily the latency requirements can be fulfilled. Moreover, the incorporation of the edge server results in an improved algorithm, decreased image-matching time and enhanced computational capability, in contrast to executing the same algorithm on the terminal side. Edge computing approaches, along with caching strategies implemented at the edge, offer a blueprint for the advancement of wireless networks beyond 5G for emerging applications, such as video streaming and high-definition images and as they have been analyzed at the network core and edge, reduce duplicity and redundant traffic. In a multi-user MEC system, it aims to reduce overall delay and energy consumption while enhancing accuracy in AI inference tasks, taking into account computation and communication resources. MEC and D2D technologies further enhance service performance by coordinating distributed resources. Additionally, a MAC scheduling scheme for multiple users is proposed for wireless VR deployments in a 5G MIMO-OFDM system, enabling maximum simultaneous VR clients while meeting demanding criteria for transfer reliability, data rate and responsiveness.

For XR services there is a strong need for a high level of security, to protect user data and fend off unauthorized access to XR applications. This includes encryption of data in transit and at rest and authentication mechanisms to ensure that only authorized users can have access. To this step, in AR/VR deployments it is used the blockchain technology to enhance security and uniqueness in (AR/VR) deployments. More specifically it is proposed a decentralized architecture that is using blockchain to support instantaneous connectivity, ample bandwidth and seamless data transfer within AR/VR applications in 5G and 6G network services. The use of blockchains and ICN is also proposed for ensuring pervasive connectivity of VR/AR within the context of 6G-enabled massive IoT. A proposed sharing architecture ICN is used to augment the effectiveness of IC-mIoT and reassure QoS for users. A new mechanism that is called PoCO is also proposed and with permissioned blockchain it is utilized to safeguard diverse forms of resource transactions and content within IC-mIoT.

The significance of cloud architecture cannot be overstated when it comes to XR/AR/VR technologies. These advancements play a crucial role in enhancing the QoS for XR-aided devices that stimulate multiple senses. Moreover,

there is a pressing need to improve the network infrastructure through the implementation of diverse caching strategies and algorithms. These techniques are specifically designed to optimize signal strength and computational load during the transition of computing tasks from MEC to 5G networks. Additionally, an intriguing concept revolves around the development of collaborative XR applications that enable users to have multi-user experiences across different platforms. This collaborative approach balances between user and service provider interests, using balance rate and mobile energy consumption as metrics. The allocation of DNN computations across the network edge, clouds server and mobile users exhibits a reduced imbalance ratio in all three situations, indicating enhanced contentment for both service providers and end-users.

Moreover, advanced QoS is indispensable for XR services. QoS acts as a distinctive feature that allows the network to prioritize XR traffic, ensuring that XR applications receive the necessary bandwidth and meet latency requirements. The synchronization of applications and coordination of QoS policies are also vital for handling multi-modal data flows among multiple UE. This coordination fosters seamless integration and smooth operation of XR applications. Furthermore, it is crucial to support differentiated QoS handling by taking into consideration the varying importance of PDUs. By treating packets belonging to less important PDU sets differently, the QoS model and policy control can be enhanced to accommodate the importance and dependency information associated with a particular PDU set. Lastly, achieving effective uplink-downlink transmission synchronization is pivotal to reach round-trip latency requisites. This coordination involves potential interaction between AF, 5G System (5GS) and potential QoS enhancements. By ensuring smooth coordination and optimization of transmission, requisites for XR applications regarding latency can be achieved effectively.

To integrate XR applications into the 5G network, it is proposed an approach based on a 5GMS model with various functions such as a 5G-XR client, a 5G-XR aware application, a 5G-XR AS, a 5G-XR Application Provider and a 5G-XR AF. The QoS model in 5G embraces Reflective QoS and employs a QFI for flow identification. They also discussed different aspects related to the evaluation of power consumption, coverage and mobility in XR, as well as the client reference model and architecture for VR QoE measurement. It also mentioned an immersive media metrics client reference model and a VR audio system for content creation and delivery. Mechanisms that enable codec/rate adaptation to meet requirements for services and 5GS information exposure for XR/media Enhancements. In this model it is used DASH over MBMS so as to expand the capabilities of broadcast networks by enabling real-time adaptation of DASH clients to changing network conditions. This ensures efficient and robust delivery of multimedia content over the broadcast network, resulting in an enhanced user experience.

Finally, they cited some matters that are related with XR improvements regarding network performance. 5G NR may be applicable in situations where the capacity in urban macro situations is lower than in dense urban and indoor hotspot scenarios, specifically for uplink video applications. The NR system's capacity is reduced for applications requiring higher data rates but increases with larger PDB values and system bandwidth. Higher PDB values result in improved mobility KPIs and lower frame rates lead to an increased number of consecutive XR packets lost. As handover interruption time increases, mobility performance declines. On the other hand, a longer handover interruption time has a detrimental effect on mobility KPIs, making them worse. Both uplink and downlink XR-Awareness contribute to enhancing gNB radio resource scheduling by considering PDU sets and Data Bursts. Various alternatives are proposed for mapping PDU sets and QoS flows in the NAS, as well as mapping QoS flows into Data Radio Bearers.

Additionally, to achieve high performance and scalability in large-scale XR systems, Raster-Based Split Rendering is commonly employed. This technique distributes the rendering workload across multiple Graphics Processing Units (GPUs), servers, or nodes. In other cases, a Spatial Computing Server works in conjunction with XR-capable devices to manage and retain information about the physical environment and the location of XR objects.

When it comes to supporting high-capacity immersive applications, scalability, security, energy-aware offloading and edge computing are crucial factors to consider. The next-generation mmWave wireless technology has the capability to address these considerations effectively. It enables the seamless expansion of networks to handle growing demands, ensures robust security measures, optimizes energy consumption through intelligent offloading and leverages edge computing capabilities for efficient processing and delivery of immersive content. Moreover, modifying Convolutional neural network (CNN) workflows for interactive broadcasting is an interesting direction [9].

AI, honeypot and SDN technologies have the potential to enhance situational awareness systems within the metaverse. These advancements can bolster monitoring capabilities for detecting large-scale distributed threats, thereby improving global situational awareness. To ensure safety and address social impact concerns, the implementation of cyber-insurance and cyber-physical social system approaches can provide significant perception into safeguarding the integrity of devices in the metaverse. Moreover, blockchain technology should be a key factor in these cases. It can establish digital identities for metaverse end-users without the need for trust and provide a possible resolution to ensure the credibility of data when creating digital twins and undertaking mitigation measures [8].

Holographic communication plays a vital role in enhancing productivity, social interaction and entertainment within immersive AR applications. The presence of lightweight AR glasses and sophisticated 3D compression algorithms has

opened up the opportunity to implement AR applications using current 5G technology. However, to fulfill the demanding prerequisites of XR in terms of latency, dependability and data bit rates, certain measures need to be taken and the recommendation is to implement effective techniques such as delay-aware scheduling and robust link adaptation in the near term [36].

Overall, XR services necessitate attentive planning, advanced network capabilities and a focus on user experience and security. By following best practices and enhancing the latest standards and technologies, network operators and service providers can successfully deploy XR services that deliver immersive and engaging experiences to users [79].

B. OPEN CHALLENGES AND FUTURE DIRECTIONS IN XR APPLICATIONS

There are several open challenges for XR applications that must be tackled in order to facilitate widespread acceptance and implementation and offer a high-quality user experience that we list below in detail:

- There are as pending, a lot of KPIs for the architectures examined before. More specifically for Viewport Rendering in network only DL Range Rate is defined. Furthermore, for Generalized Split Rendering and XR Distributed Computing all KPI features are under discussion for further study. For XR Conversational and Raster-based Split Rendering with Pose Correction, Traffic file size distribution needs to be reexamined in the future. For the latter architecture Traffic periodicity range is still unknown. Finally, for XR Conferencing DL and UL per range are matters for additional future study. In a research paper [80], a new and innovative cross-layer Medium Access Control (MAC) scheduling method named "Application-aware MAC" is introduced. This technique greatly improves network efficiency by launching a slight deviation in predictions at the application layer. Additionally, an altered split rendering architecture is suggested, which integrates additional Network Function Virtualizations (NFVs) to simplify the implementation of this concept.
- XR and VR Gaming application applications demand significant amount of network capacity to transfer substantial data flows in real-time. This places a tension on network infrastructure and can result in network congestion, latency and reliability cases. New network techniques, such as 5G and network slicing, are being developed to address these challenges and meet the expected traffic growth [81].
- Nowadays there are different XR platforms, devices and formats, that could make it triggering to create content that works across all platforms. Interoperability standards are crucial to enable seamless content creation and distribution across different platforms. A scalable architecture for XR applications should be bolstered by forthcoming progress in network technologies, such as edge computing and 5G [82].
- The design of applications must take into account human factors such as visual awareness, depth perception and the user's engagement to collaborate with virtual objects. For that purpose the design of XR applications must consider the physiological and psychological effects of extended exposure to virtual environments [83].
- Researchers are following different strategies to reduce or eliminate motion sickness, such as adjusting the FOV or using haptic feedback. Many users face this problem when using VR or XR applications, which can limit the duration and quality of their experience [84].
- Upgrade in hardware technology is needed to offer a more immersive and high-end user experience as the current state of hardware technology for AR, VR, or XR applications is limited with respect to FOV, resolution and battery life [85].
- Distributing XR content to users in a timely and efficient way is a noteworthy challenge, especially for large or complex content such as video or 3D models. New distribution technologies, such as peer-to-peer networks or content delivery networks (CDNs) should help address this challenge [86].
- With the increasing expansion of online interactive network services, it is possible to employ online IE evaluation methods as a potential future option. It is important for upcoming studies to assess how the introduction of more intricate applications affects user's overall QoE in 5G networks.
- In the future, it will be crucial to develop reliable decentralized applications for AR/VR that are built on the 6G network. Furthermore, there is a requirement to explore the effects of incorporating blockchain-based trusted services on the 6G network, specifically for the widespread adoption of Internet of Things (IoT) deployments.
- Looking ahead, MEC will be instrumental in enabling real-time processing tasks and as method to bring applications closer to data sources. It offers numerous opportunities for research and development, including optimizing network resource utilization, enhancing application performance and reducing latency. In addition, MEC will be instrumental in supporting AR/VR applications and autonomous applications. Caching and computing at the edge will play a significant role in ensuring consistent performance across different applications. By leveraging BSs and employing slicing operations, MEC models can effectively achieve low latency, although the computing power of the BS will impact the quality of visuals.
- In upcoming days, there will be a growing interest in collaborative AR experiments that involve multiple AR/VR devices. These experiments aim to analyze the

system-level performance and optimize it accordingly. The findings indicate that collaborative solutions outperform self-contained cloud or edge-based offloading methods when it comes to AR applications in 5G networks. Furthermore, there is a need to explore location-based services and potential advancements for enhancing mobile Web AR. This includes On-Web AI Service, multi-edge collaboration AI, 5G-Enabled Networking AI and network slicing, all of which hold promising prospects for the future.

- In the future there is a need to design algorithms that analyze various forms of interference within 5G networks and assess the tradeoff between handoff rates and network interference. To achieve this, there is a requirement for enhancements that can support both unicast and broadcast streams simultaneously, while also incorporating beamforming for broadcast applications. Additionally, it is necessary to reevaluate the tradeoff between improved spectral efficiency and the associated increase in BS costs, considering the implications more extensively.
- For future advancements in 5G, there are several potential improvements that can enhance the user experience. These include traffic-awareness, network coding, interference coordination, semi-persistent scheduling and mobility-related enhancements. Specifically, effective scheduling strategies for XR traffic streams are proposed. These strategies aim to enable XR applications to coexist with other applications within the network and to validate models for higher FPS values, ensuring their effectiveness in practical scenarios.
- Potential solutions for addressing packet jitter are being explored, including the extension of PDCCH monitoring duration. While this approach increases power consumption, it has the potential to effectively mitigate jitter. In addition, future prospects should focus on investigating power-saving enhancements and synchronization of uplink and downlink transmissions, along with the reporting of assistance information by UE. These improvements have the potential to further optimize power consumption and enhance overall system performance.
- Suggested architectures have practical implications for the widespread implementation of wireless XR in B5G systems. Future endeavors should focus on developing improved scheduling algorithms to further enhance the efficiency of these architectures. Furthermore, the feasibility of ubiquitous wireless VR applications can be achieved through cross-system strategy and algorithm design. However, there is a need for future research to tackle security topics within the suggested framework.
- Further exploration is needed in compression and encoding techniques to attain minimal latency, interactivity and also improved efficiency. Moreover, it is important to develop techniques, such as cloud-based streaming, that can enhance projection processing and allocate a higher number of pixels specifically to the user's viewport. These advancements will contribute to a more seamless and immersive experience by reducing latency and maximizing the utilization of available resources [9].
- The integration of advanced technologies in data management introduces more vulnerabilities that can be targeted for potential attacks. In traditional online services, platform operators have control over user data, often utilizing it for their own profit. This centralized approach poses privacy risks, increasing the possibility of data misuse. It is imperative for users to regain control over their personal information to mitigate these concerns. It is important to develop and enforce security mechanisms to reduce and eliminate possible dangers and risks [8].
- Improved QoS in 5G networks is more than important to encourage the broad implementation of XR services. 5G QoS framework and design enhancements will be implemented in order to meet future XR QoS requirements (Quality of Service management methods throughout various levels of the 5G NR protocol architecture, encompassing both the radio access and core networks). Also, architectures for handling XR's unique characteristics and presenting adaptation mechanisms for extended reality traffic in the application layer and additionally, MAC layer QoS schedulers designed to maximize performance across multiple protocol layers [87]. Additionally, in [88] it is presented a comprehensive evaluation of QoS schedulers for extended reality traffic in various scenarios in NR 5G system. It is introduced and assessed a novel QoS MAC scheduler that exploits the QoS model implemented in 5G NR, specifically designed to enhance XR applications. The evaluation process begins by verifying the correct functioning of the QoS MAC scheduler in a single-cell environment. Both saturation and non-saturation scenarios are examined to ensure its proper operation. Furthermore, the evaluation extends to a practical multi-cell scenario specified by International Telecommunication Union - Radiocommunication Sector (ITU-R). With the objective of demonstrating its effectiveness, the suggested QoS MAC scheduler showcases its capability to accurately organize XR traffic and other types of traffic according to the requested QoS indicators.
- Upcoming Release 18 in 3GPP is poised to deliver substantial improvements to the capabilities of 5G and XR, focusing on improving performance, enabling more flexible spectrum utilization, supporting diverse devices and evolving network topology for various deployments. This release will also incorporate data-driven, intelligent network solutions by leveraging AI and ML technologies. Moreover, as 6G standardization

is anticipated to commence around 2025 within 3GPP, Release 18 will act as a foundation for the advancement of this upcoming era of wireless technology. To enhance 5G performance, 3GPP will continue its efforts in Release 18 by exploring areas such as conserving energy within the network, extending coverage, facilitating mobile connections, advancing MIMO technologies, implementing the Multiple Beam System (MBS) and improving positioning capabilities. While positioning support was initially introduced in Release 15 NR, further improvements were made in Release 16 NR, expanding the range of positioning methods and enhancing accuracy and latency for critical use cases like remote control and factory automation. Building on these advancements, Release 17 NR introduced further improvements to decrease latency further, achieve higher positioning accuracy (20-30 cm) for specific use cases and enhance the protection of location information. Looking ahead, 3GPP Release 18 aims to explore options for enhancing precision, reliability and energy efficiency in positioning techniques. It will also delve into sidelink positioning and explore positioning support for RedCap devices [89]. Finally, in 6G networks, critical services will rely on edge computing and programmable infrastructure, enabling E2E network slicing for assured QoS. Interactive services in real-world domains require efficient data movement for distributed decision making. Integrating application and network layers enables semantic approaches, increasing efficiency, reducing energy consumption and improving user QoS [12].

V. CONCLUSION

This paper provides a comprehensive examination of enabling XR services in beyond 5G networks. Specifically, we have presented a thorough survey and taxonomy of scientific papers in this topic highlighting the different solutions to different parts of the network. Afterwards, we focus on the progress within 3GPP, where releases from 17 and 18 have been studied in order to summarize the corresponding solutions, KPIs, services and networking element architectures. Finally, lessons learned and open challenges are provided as a summary of our survey and the road ahead towards mobile immersive experiences in future 6G networks. We anticipate that this survey can provide insights into the mobile XR services landscape as the main enabler towards efficient immersive experience services in 6G networks.

REFERENCES

[1] F. Tang, X. Chen, M. Zhao, and N. Kato, "The roadmap of communication and networking in 6G for the metaverse," *IEEE Wireless Commun.*, vol. 30, no. 4, pp. 72–81, Aug. 2024, doi: 10.1109/MWC.019.2100721.

[2] A. M. Aslam, R. Chaudhary, A. Bhardwaj, I. Budhiraja, N. Kumar, and S. Zeadally, "Metaverse for 6G and beyond: The next revolution and deployment challenges," *IEEE Internet Things Mag.*, vol. 6, no. 1, pp. 32–39, Mar. 2023, doi: 10.1109/IOTM.001.2200248.

[3] H. Peng, P.-C. Chen, P.-H. Chen, Y.-S. Yang, C.-C. Hsia, and L.-C. Wang, "6G toward metaverse: Technologies, applications, and challenges," in *Proc. IEEE VTS Asia–Pacific Wireless Commun. Symp. (APWCS)*, Aug. 2022, pp. 6–10, doi: 10.1109/APWCS55727.2022.9906483.

[4] L. Chang, Z. Zhang, P. Li, S. Xi, W. Guo, Y. Shen, Z. Xiong, J. Kang, D. Niyato, X. Qiao, and Y. Wu, "6G-enabled edge AI for metaverse: Challenges, methods, and future research directions," *J. Commun. Inf. Netw.*, vol. 7, no. 2, pp. 107–121, Jun. 2022.

[5] T. Taleb, A. Boudi, L. Rosa, L. Cordeiro, T. Theodoropoulos, K. Tserpes, P. Dazzi, A. I. Protopsaltis, and R. Li, "Toward supporting XR services: Architecture and enablers," *IEEE Internet Things J.*, vol. 10, no. 4, pp. 3567–3586, Feb. 2023.

[6] F. Hu, Y. Deng, H. Zhou, T. H. Jung, C.-B. Chae, and A. H. Aghvami, "A vision of an XR-aided teleoperation system toward 5G/B5G," *IEEE Commun. Mag.*, vol. 59, no. 1, pp. 34–40, Jan. 2021, doi: 10.1109/MCOM.001.2000581.

[7] G. Minopoulos and K. E. Psannis, "Opportunities and challenges of tangible XR applications for 5G networks and beyond," *IEEE Consum. Electron. Mag.*, early access, Mar. 22, 2022, doi: 10.1109/MCE.2022.3156305.

[8] Y. Wang, Z. Su, N. Zhang, R. Xing, D. Liu, T. H. Luan, and X. Shen, "A survey on metaverse: Fundamentals, security, and privacy," *IEEE Commun. Surveys Tuts.*, vol. 25, no. 1, pp. 319–352, 1st Quart., 2023, doi: 10.1109/COMST.2022.3202047.

[9] A. Yaqoob, T. Bi, and G.-M. Muntean, "A survey on adaptive 360° video streaming: Solutions, challenges and opportunities," *IEEE Commun. Surveys Tuts.*, vol. 22, no. 4, pp. 2801–2838, 4th Quart., 2020, doi: 10.1109/COMST.2020.3006999.

[10] P. Hande, P. Tinnakornsrisuphap, J. Damnjanovic, H. Xu, M. Mondet, H. Y. Lee, and I. Sakhnini, "Extended reality over 5G—Standards evolution," *IEEE J. Sel. Areas Commun.*, vol. 41, no. 6, pp. 1757–1771, Jun. 2023.

[11] M. Gapeyenko, V. Petrov, S. Paris, A. Marcano, and K. I. Pedersen, "Standardization of extended reality (XR) over 5G and 5G-advanced 3GPP new radio," *IEEE Network*, vol. 37, no. 4, pp. 22–28, Jul./Aug. 2023, doi: 10.1109/MNET.003.2300062.

[12] W. Chen, X. Lin, J. Lee, A. Toskala, S. Sun, C. F. Chiasserini, and L. Liu, "5G-advanced toward 6G: Past, present, and future," *IEEE J. Sel. Areas Commun.*, vol. 41, no. 6, pp. 1592–1619, Jun. 2023, doi: 10.1109/JSAC.2023.3274037.

[13] Y. Gao, X. Wei, J. Chen, and L. Zhou, "Toward immersive experience: Evaluation for interactive network services," *IEEE Netw.*, vol. 36, no. 1, pp. 144–150, Jan. 2022, doi: 10.1109/MNET.121.2100323.

[14] P. Bhattacharya, D. Saraswat, A. Dave, M. Acharya, S. Tanwar, G. Sharma, and I. E. Davidson, "Coalition of 6G and blockchain in AR/VR space: Challenges and future directions," *IEEE Access*, vol. 9, pp. 168455–168484, 2021, doi: 10.1109/ACCESS.2021.3136860.

[15] S. Liao, J. Wu, J. Li, and K. Konstantin, "Information-centric massive IoT-based ubiquitous connected VR/AR in 6G: A proposed caching consensus approach," *IEEE Internet Things J.*, vol. 8, no. 7, pp. 5172–5184, Apr. 2021, doi: 10.1109/JIOT.2020.3030718.

[16] A. Vidal-Balea, O. Blanco-Novoa, P. Fraga-Lamas, and T. M. Fernandez-Carames, "A multi-platform collaborative architecture for multi-user extended reality applications," in *Proc. 5th XoveTIC Conf.*, 2023, pp. 148–151.

[17] S. Verde, M. Marcon, S. Milani, and S. Tubaro, "Advanced assistive maintenance based on augmented reality and 5G networking," *Sensors*, vol. 20, no. 24, p. 7157, 2020, doi: 10.3390/s20247157.

[18] J. Cao, X. Liu, X. Su, S. Tarkoma, and P. Hui, "Context-aware augmented reality with 5G edge," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, Dec. 2021, pp. 1–6, doi: 10.1109/globecom46510.2021.9685498.

[19] X. Qiao, P. Ren, S. Dustdar, and J. Chen, "A new era for web AR with mobile edge computing," *IEEE Internet Comput.*, vol. 22, no. 4, pp. 46–55, Jul. 2018, doi: 10.1109/MIC.2018.043051464.

[20] Y. Wang, T. Yu, and K. Sakaguchi, "Context-based MEC platform for augmented-reality services in 5G networks," in *Proc. IEEE 94th Veh. Technol. Conf. (VTC-Fall)*, Norman, OK, USA, Sep. 2021, pp. 1–5, doi: 10.1109/VTC2021-Fall52928.2021.9625304.

[21] S. Sukhmani, M. Sadeghi, M. Erol-Kantarci, and A. El Saddik, "Edge caching and computing in 5G for mobile AR/VR and tactile internet," *IEEE MultimediaMag.*, vol. 26, no. 1, pp. 21–30, Jan. 2019, doi: 10.1109/MMUL.2018.2879591.

[22] P. Zhou, S. Fu, B. Finley, X. Li, S. Tarkoma, J. Kangasharju, M. Ammar, and P. Hui, "5G MEC computation handoff for mobile augmented reality," 2021, *arXiv:2101.00256*.

[23] H. Zhu, Y. Li, Z. Chen, and L. Song, "Mobile edge resource optimization for multiplayer interactive virtual reality game," in *Proc. IEEE Wireless Commun. Netw. Conf. (WCNC)*, Nanjing, China, Mar. 2021, pp. 1–6, doi: 10.1109/WCNC49053.2021.9417124.

[24] A. Casparsen, F. Chiariotti, and J. J. Nielsen, "On-the-fly edge transcoding for interactive VR," in *Proc. IEEE 20th Consum. Commun. Netw. Conf. (CCNC)*, Jan. 2023, pp. 863–866, doi: 10.1109/ccnc51644.2023.10060308.

[25] G. Pan, H. Zhang, S. Xu, S. Zhang, and X. Chen, "Joint optimization of video-based AI inference tasks in MEC-assisted augmented reality systems," *IEEE Trans. Cognit. Commun. Netw.*, vol. 9, no. 2, pp. 479–493, Apr. 2023, doi: 10.1109/TCCN.2023.3235773.

[26] X. Liu and Y. Deng, "Learning-based prediction, rendering and association optimization for MEC-enabled wireless virtual reality (VR) networks," *IEEE Trans. Wireless Commun.*, vol. 20, no. 10, pp. 6356–6370, Oct. 2021, doi: 10.1109/TWC.2021.3073623.

[27] Z. Chen, H. Zhu, L. Song, D. He, and B. Xia, "Wireless multiplayer interactive virtual reality game systems with edge computing: Modeling and optimization," *IEEE Trans. Wireless Commun.*, vol. 21, no. 11, pp. 9684–9699, Nov. 2022, doi: 10.1109/TWC.2022.3178618.

[28] M. Liubogoshchev, K. Ragimova, A. Lyakhov, S. Tang, and E. Khorov, "Adaptive cloud-based extended reality: Modeling and optimization," *IEEE Access*, vol. 9, pp. 35287–35299, 2021, doi: 10.1109/ACCESS.2021.3062555.

[29] A. Mihaljevic, A. Keselj, and A. Lipovac, "Impact of 5G network performance on augmented reality application QoE," in *Proc. Int. Conf. Softw., Telecommun. Comput. Netw. (SoftCOM)*, Sep. 2021, pp. 1–4, doi: 10.23919/SoftCOM52868.2021.9559067.

[30] B. Krogfoss, J. Duran, P. Perez, and J. Bouwen, "Quantifying the value of 5G and edge cloud on QoE for AR/VR," in *Proc. 12th Int. Conf. Quality Multimedia Exper. (QoMEX)*, 2020, pp. 1–4.

[31] P. Lin, Q. Song, F. R. Yu, D. Wang, A. Jamalipour, and L. Guo, "Wireless virtual reality in beyond 5G systems with the Internet of Intelligence," *IEEE Wireless Commun.*, vol. 28, no. 2, pp. 70–77, Apr. 2021, doi: 10.1109/MWC.001.2000303.

[32] Z. Nadir, T. Taleb, H. Flinck, O. Bouachir, and M. Bagaa, "Immersive services over 5G and beyond mobile systems," *IEEE Netw.*, vol. 35, no. 6, pp. 299–306, Nov. 2021, doi: 10.1109/MNET.121.2100172.

[33] P. Ren, X. Qiao, Y. Huang, L. Liu, S. Dustdar, and J. Chen, "Edge-assisted distributed DNN collaborative computing approach for mobile web augmented reality in 5G networks," *IEEE Netw.*, vol. 34, no. 2, pp. 254–261, Mar. 2020, doi: 10.1109/MNET.011.1900305.

[34] P. Ren, X. Qiao, Y. Huang, L. Liu, C. Pu, S. Dustdar, and J. Chen, "Edge AR x5: An edge-assisted multi-user collaborative framework for mobile web augmented reality in 5G and beyond," *IEEE Trans. Cloud Comput.*, vol. 10, no. 4, pp. 2521–2537, Oct. 2022, doi: 10.1109/TCC.2020.3046128.

[35] P. Ren, X. Qiao, Y. Huang, L. Liu, C. Pu, and S. Dustdar, "Fine-grained elastic partitioning for distributed DNN towards mobile web AR services in the 5G era," *IEEE Trans. Services Comput.*, vol. 15, no. 6, pp. 3260–3274, Nov. 2022, doi: 10.1109/TSC.2021.3098816.

[36] Spotlight on Extended Reality. (2021). *Charting the future of Innovation, Volume 110, Ericsson*. [Online]. Available: https://www.ericsson.com/en/reports-and-papers/ericssontechnology-review/articles/spotlight-on-xr

[37] J. Wallace and A. Valdivia, "A high-performance 5G/6G infrastructure for augmented, virtual and extended reality," in *Proc. Int. Conf. Comput. Sci. Comput. Intell. (CSCI)*, 2021, pp. 1291–1296, doi: 10.1109/CSCI54926.2021.00264.

[38] X. Ge, L. Pan, Q. Li, G. Mao, and S. Tu, "Multipath cooperative communications networks for augmented and virtual reality transmission," *IEEE Trans. Multimedia*, vol. 19, no. 10, pp. 2345–2358, Oct. 2017, doi: 10.1109/TMM.2017.2733461.

[39] A. Prasad, M. A. Uusitalo, D. Navratil, and M. Saily, "Challenges for enabling virtual reality broadcast using 5G small cell network," in *Proc. IEEE Wireless Commun. Netw. Conf. Workshops (WCNCW)*, Barcelona, Spain, Apr. 2018, pp. 220–225, doi: 10.1109/WCNCW.2018.8368976.

[40] J. K. Sundararajan, H. J. Kwon, O. Awoniyi-Oteri, Y. Kim, C. P. Li, J. Damnjanovic, S. Zhou, R. Ma, Y. Tokgoz, P. Hande, and T. Luo, "Performance evaluation of extended reality applications in 5G NR system," in *Proc. IEEE 32nd Annu. Int. Symp. Personal, Indoor Mobile Radio Commun. (PIMRC)*, Helsinki, Finland, Sep. 2021, pp. 1–7, doi: 10.1109/PIMRC50174.2021.9569585.

[41] D. Li, H. You, W. Jiang, X. Chen, C. Zeng, and X. Sun, "Enhanced power saving schemes for eXtended reality," in *Proc. IEEE 32nd Annu. Int. Symp. Pers., Indoor Mobile Radio Commun. (PIMRC)*, Helsinki, Finland, Sep. 2021, pp. 1–6, doi: 10.1109/PIMRC50174.2021.9569348.

[42] Y. Kim, H.-J. Kwon, O. Awoniyi-Oteri, P. Hande, J. K. Sundararajan, Y. Tokgoz, T. Luo, K. Mukkavilli, and T. Ji, "UE power saving techniques for extended reality (XR) services in 5G NR systems," in *Proc. IEEE 32nd Annu. Int. Symp. Pers., Indoor Mobile Radio Commun. (PIMRC)*, Helsinki, Finland, Sep. 2021, pp. 1–7, doi: 10.1109/PIMRC50174.2021.9569722.

[43] S. Paris, K. Pedersen, and Q. Zhao, "Adaptive discontinuous reception in 5G advanced for extended reality applications," in *Proc. IEEE 95th Veh. Technol. Conf.*, Helsinki, Finland, Jun. 2022, pp. 1–6, doi: 10.1109/VTC2022-Spring54318.2022.9860663.

[44] M. Lecci, M. Drago, A. Zanella, and M. Zorzi, "An open framework for analyzing and modeling XR network traffic," *IEEE Access*, vol. 9, pp. 129782–129795, 2021, doi: 10.1109/ACCESS.2021.3113162.

[45] Y. Gao, S. Xue, M. Ding, J. Peng, and J. Pang, "Exploring extended reality with flexible spectrum access in wireless cellular network," in *Proc. IEEE 32nd Annu. Int. Symp. Personal, Indoor Mobile Radio Commun. (PIMRC)*, Helsinki, Finland, 2021, pp. 1–6, doi: 10.1109/PIMRC50174.2021.9569526.

[46] P. Paymard, A. Amiri, T. E. Kolding, and K. I. Pedersen, "Enhanced link adaptation for extended reality code block group based HARQ transmissions," in *Proc. IEEE Globecom Workshops*, Dec. 2022, pp. 711–716.

[47] P. Paymard, A. Amiri, T. E. Kolding, and K. I. Pedersen, "Extended reality over 3GPP 5G-advanced new radio: Link adaptation enhancements," 2022, *arXiv:2210.14578*.

[48] D. Gonzalez Morin, D. Medda, A. Iossifides, P. Chatzimisios, A. Garcia Armada, A. Villegas, and P. Perez, "An eXtended reality offloading IP traffic dataset and models," 2023, *arXiv:2301.11217*.

[49] Z. Huang, C. Xiong, H. Ni, D. Wang, Y. Tao, and T. Sun, "Standard evolution of 5G-advanced and future mobile network for extended reality and metaverse," *IEEE Internet Things Mag.*, vol. 6, no. 1, pp. 20–25, Mar. 2023, doi: 10.1109/IOTM.001.2200261.

[50] J. Dai, G. Yue, S. Mao, and D. Liu, "Sidelink-aided multiquality tiled 360° virtual reality video multicast," *IEEE Internet Things J.*, vol. 9, no. 6, pp. 4584–4597, Mar. 2022, doi: 10.1109/JIOT.2021.3105100.

[51] X. Liu, X. Li, and Y. Deng, "Learning-based prediction and proactive uplink retransmission for wireless virtual reality network," *IEEE Trans. Veh. Technol.*, vol. 70, no. 10, pp. 10723–10734, Oct. 2021, doi: 10.1109/TVT.2021.3102844.

[52] M. Huang and X. Zhang, "MAC scheduling for multiuser wireless virtual reality in 5G MIMO-OFDM systems," in *Proc. IEEE Int. Conf. Commun. Workshops (ICC Workshops)*, Kansas City, MO, USA, May 2018, pp. 1–6, doi: 10.1109/ICCW.2018.8403486.

[53] B. Bojovic, S. Lagén, K. Koutlia, X. Zhang, P. Wang, and L. Yu, "Enhancing 5G QoS management for XR traffic through XR loopback mechanism," *IEEE J. Sel. Areas Commun.*, vol. 41, no. 6, pp. 1772–1786, Jun. 2023, doi: 10.1109/JSAC.2023.3273701.

[54] *Extended Reality (XR) in 5G; 3rd Generation Partnership Project; Technical Specification Group Services and System Aspects; Extended Reality (XR) in 5G (Release 18)*, document TR 26.928 V18.0.0, 3GPP, Mar. 2023.

[55] *Study on XR Enhancements for NR; 3rd Generation Partnership Project; Technical Specification Group Services and System Aspects (Release 18)*, document TR 38.835 V0.3.1, 3GPP, 2022.

[56] *Traffic Models and Quality Evaluation Methods for Media and XR Services in 5G Systems; 3rd Generation Partnership Project; Technical Specification Group Services and System Aspects (Release 18)*, document TR Gl V1.2.0, 3GPP, 2022.

[57] *Study on XR (Extended Reality) Evaluations for NR; 3rd Generation Partnership Project; Technical Specification Group Radio Access Network (Release 17)*, document TR 38.838, V17.0.0, 3GPP, Dec. 2021.

[58] *QoE Parameters and Metrics Relevant to the Virtual Reality (VR) User Experience; 3rd Generation Partnership Project; Technical Specification Group Services and System Aspects (Release 17)*, document TR 26.929, V17.0.0, 3GPP, Apr. 2022.

[59] *Virtual Reality (VR) Media Services Over 3GPP; 3rd Generation Partnership Project; Technical Specification Group Services and System Aspects (Release 17)*, document TR 26.918, V17.0.0, 3GPP, Apr. 2022.

[60] *Virtual Reality (VR) Profiles for Streaming Applications; 3rd Generation Partnership Project; Technical Specification Group Services and System Aspects (Release 17)*, document TS 26.118, V18.0.0, 3GPP, Mar. 2023.
[61] *Support of 5G Glass-Type Augmented Reality/Mixed Reality (AR/MR) Devices; 3rd Generation Partnership Project; Technical Specification Group Services and System Aspects (Release 17)*, document TR 26.998, V18.1.0, 3GPP, Mar. 2024.
[62] *Study on XR (Extended Reality) and Media Services; 3rd Generation Partnership Project; Technical Specification Group Services and System Aspects (Release 18)*, document TR 23.700-60, V1.3.0, 3GPP, Nov. 2022.
[63] Extended Reality and 3GPP Evolutions, "A 5G americas white paper," 5G Americas, Bellevue, WA, USA, Nov. 2022.
[64] *Traffic Models and Quality Evaluation Methods for Media and XR Services in 5G Systems; 3rd Generation Partnership Project; Technical Specification Group Services and System Aspects (Release 18)*, document TR 26.926, V1.2.0, Aug. 2022.
[65] I. F. Akyildiz and H. Guo, "Wireless communication research challenges for extended reality (XR)," *ITU J. Future Evolving Technol.*, vol. 3, no. 2, pp. 273–287, Apr. 2022.
[66] *XR Use Cases, Evaluation Methodologies and Traffic Model CATT, 3GPP TSG RAN WG1 Meeting #103-e*, document R1-2007843, Oct. 2020.
[67] *Discussion on Applications and Evaluation Methodology for XR Services*, document R1-2009041, Xiaomi, 3GPP TSG RAN WG1 Meeting #103-e, Oct. 2020.
[68] S. Ahsan, S. Mate, I. D. D. Curcio, A. Aminlou, Y. You, E. B. Aksu, and M. M. Hannuksela, "Viewport-dependent delivery for conversational immersive video," *IEEE Access*, vol. 10, pp. 129539–129551, 2022, doi: 10.1109/ACCESS.2022.3225231.
[69] *TSG-SA4 Meeting*, document 112-e. S4-210124, 3GPP, Feb. 2021.
[70] *System Architecture for the 5G System; 3rd Generation Partnership Project; Technical Specification Group Services and System Aspects (5GS) Stage 2 (Release 18)*, document TS 23.501, V18.5.0, 3GPP, Mar. 2024.
[71] *Transparent End-to-end Packet-switched Streaming Service (PSS); Progressive Download and Dynamic Adaptive Streaming Over HTTP (3GP-DASH) 3rd Generation Partnership Project; Technical Specification Group Services and System Aspects (Release 18)*, document TS 26.247, V18.0.0, 3GPP, Mar. 2024.
[72] *Interface Between the Control Plane and the User Plane Nodes;Stage 3 3rd Generation Partnership Project; Technical Specification Group Services and System Aspects (Release 18)*, document TS 29.244, V18.2.1, 3GPP, Jun. 2023.
[73] *Network Data Analytics Services; 3rd Generation Partnership Project; Technical Specification Group Services and System Aspects (Release 15)*, document TS 29.520, V 18.5.1, 3GPP, Apr. 2024.
[74] S. Dutta, D. Roy, and G. Das, "XR-specific C-DRX enhancement for UE power saving in 5G NR," in *Proc. IEEE 33rd Annu. Int. Symp. Pers., Indoor Mobile Radio Commun. (PIMRC)*, Kyoto, Japan, Sep. 2022, pp. 1–6, doi: 10.1109/PIMRC54779.2022.9977925.
[75] *Study on XR Enhancements for NR; 3rd Generation Partnership Project; Technical Specification Group Services and System Aspects (Release 18)*, document TR 38.835, V0.3.1, 3GPP, Nov. 2022.
[76] *Study on Scenarios and Requirements for Next Generation Access Technologies; 3rd Generation Partnership Project; Technical Specification Group Services and System Aspects (Release 17)*, document TR 38.913, V 17.0.0, 3GPP, Mar. 2022.
[77] *Study on Management and Orchestration of Network Slicing for Next Generation Network; 3rd Generation Partnership Project; Technical Specification Group Services and System Aspects(Release 15)*, document 28.801, 3GPP, Jan. 2018.
[78] Open Networking Foundation. (2017). *Mobile Edge Computing: Unleashing the Power of the Network Edge*. [Online]. Available: https://www.opennetworking.org/images/stories/downloads/sdn-resources/technical-reports/TR-MEC-White-Paper.pdf
[79] S. N. B. Gunkel, E. Potetsianakis, T. E. Klunder, A. Toet, and S. S. Dijkstra-Soudarissanane, "Immersive experiences and XR: A game engine or multimedia streaming problem?" *SMPTE Motion Imag. J.*, vol. 132, no. 5, pp. 30–37, Jun. 2023, doi: 10.5594/JMI.2023.3269752.
[80] S. Dutta, D. Roy, and G. Das, "Modified split-rendering architecture to enable AI-assisted application-aware MAC for XR slice," *IEEE Netw. Lett.*, early access, Jun. 7, 2023, doi: 10.1109/LNET.2023.3283701.
[81] T. Iwai and A. Nakao, "Sliceable congestion control for latency-awarebandwidth allocation in network slicing," in *Proc. IEEE Int. Conf. Commun. (ICC)*, Montreal, QC, Canada, 2021, pp. 1–6, doi: 10.1109/ICC42927.2021.9500605.
[82] L. Lin, X. Liao, H. Jin, and P. Li, "Computation offloading toward edge computing," *Proc. IEEE*, vol. 107, no. 8, pp. 1584–1607, Aug. 2019, doi: 10.1109/JPROC.2019.2922285.
[83] J. L. Higuera-Trujillo, J. L.-T. Maldonado, and C. L. Millán, "Psychological and physiological human responses to simulated and real environments: A comparison between photographs, 360° panoramas, and virtual reality," *Appl. Ergonom.*, vol. 65, pp. 398–409, 2017, doi: 10.1016/j.apergo.2017.05.006.
[84] S.-H. Liu, N.-H. Yu, L. Chan, Y.-H. Peng, W.-Z. Sun and M. Y. Chen, "PhantomLegs: Reducing virtual reality sickness using head-worn haptic devices," in *Proc. IEEE Conf. Virtual Reality 3D User Interfaces (VR)*, Osaka, Japan, 2019, pp. 817–826, doi: 10.1109/VR.2019.8798158.
[85] R. Mohedano and J. Chaves, "Visual interfaces in XR," in *Roadmapping Extended Reality: Fundamentals and Applications*, 2022, pp. 103–133.
[86] G. Carofiglio, G. Morabito, L. Muscariello, I. Solis, and M. Varvello, "From content delivery today to information centric networking," *Comput. Netw., Int. J. Comput. Telecommun. Netw.*, vol. 57, pp. 3116–3127, 2013, doi: 10.1016/j.comnet.2013.07.002.
[87] S. Lagen, B. Bojovic, K. Koutlia, X. Zhang, P. Wang, and Q. Qu, "QoS management for XR traffic in 5G NR: A multi-layer system view & end-to-end evaluation," *IEEE Commun. Mag.*, vol. 61, no. 12, pp. 192–198, Dec. 2023, doi: 10.1109/MCOM.015.2200745.
[88] K. Koutlia, B. Bojovic, S. Lagén, X. Zhang, P. Wang, and J. Liu, "System analysis of QoS schedulers for XR traffic in 5G NR," *Simul. Model. Pract. Theory*, vol. 125, May 2023, Art. no. 102745, doi: 10.1016/j.simpat.2023.102745.
[89] X. Lin, "An overview of 5G advanced evolution in 3GPP release 18," *IEEE Commun. Standards Mag.*, vol. 6, no. 3, pp. 77–83, Sep. 2022, doi: 10.1109/MCOMSTD.0001.2200001.

E. STAFIDAS received the B.Sc. and M.Sc. degrees in communication systems and networks from the Department of Informatics and Telecommunications, University of Athens, in 2007 and 2009, respectively. He is currently pursuing the Ph.D. degree with the Department of Informatics and Telecommunication, University of Thessaly, with a focus on beyond 5G mobile networking architectures for immersive experience services. His research interests include mobile networking protocols and architectures.

F. FOUKALAS (Senior Member, IEEE) received the Diploma degree in electrical and computer engineering from the Aristotle University of Thessaloniki, in 2001, the M.Sc. degree in technology systems from the National Technical University of Athens, in 2004, and the Ph.D. degree in informatics and telecommunications from the National and Kapodistrian University of Athens, in 2011. He is currently with the University of Thessaly, as an Assistant Professor of wireless mobile networks. His core expertise is on signal and information processing for communications, networking, and computing systems, where he has published a number of IEEE journals and conferences.