Received 17 May 2023; accepted 8 June 2023. Date of publication 19 June 2023; date of current version 13 July 2023.

Digital Object Identifier 10.1109/OJCOMS.2023.3286473

Enhancing Next-Generation Extended Reality Applications With Coded Caching

MOHAMMADJAVAD SALEHI[1] (Member, IEEE), KARI HOOLI[2], JARI HULKKONEN[2], AND ANTTI TÖLLI[1] (Senior Member, IEEE)

[1]Center for Wireless Communications, University of Oulu, 90014 Oulu, Finland

[2]Radio Research, Nokia Standards, 90620 Oulu, Finland

CORRESPONDING AUTHOR: M. SALEHI (e-mail: mohammadjavad.salehi@oulu.fi)

This work was supported in part by the Academy of Finland under Grant 318927 (6Genesis Flagship) and Grant 343586 (CAMAIDE), and in part by the Finnish Research Impact Foundation under Project 3D-WIDE.

ABSTRACT The next evolutionary step in human-computer interfaces will bring forward immersive digital experiences that submerge users in a 3D world while allowing them to interact with virtual or twin objects. Accordingly, various collaborative extended reality (XR) applications are expected to emerge, imposing stringent performance requirements on the underlying wireless connectivity infrastructure. In this paper, we examine how novel multi-antenna coded caching (CC) techniques can facilitate high-rate low-latency communications and improve users' quality of experience (QoE) in our envisioned multi-user XR scenario. Specifically, we discuss how these techniques make it possible to prioritize the content relevant to wireless bottleneck areas while enabling the cumulative cache memory of the users to be utilized as an additional communication resource. In this regard, we first explore recent advancements in multi-antenna CC that facilitate the efficient use of distributed in-device memory resources. Then, we review how XR application requirements are addressed within the third-generation partnership project (3GPP) framework and how our envisioned XR scenario relates to the foreseen use cases. Finally, we identify new challenges arising from integrating CC techniques into multi-user XR scenarios and propose novel solutions to address them in practice.

INDEX TERMS 3GPP, coded caching, extended reality, multi-antenna communications, standardization.

I. INTRODUCTION

AN INCREASINGLY large share of today's population is continuously connected to the virtual information world through flat-screen mobile devices, such as smartphones and tablets. The next step in the evolution of human-computer interfaces will bring forward immersive viewing experiences that submerge users into the 3D digital world with six degrees of freedom (forward/back, up/down, left/right, yaw, pitch, roll), thus allowing them to interact with different virtual objects while remaining integrated into the real world. Although the idea of digital immersion has been around for decades, its mass adoption has been severely hampered by movement-restricting wired connections between the interfacing headsets and the external hardware. However, this trend is recently starting to change as a new generation of powerful and affordable untethered headsets is introduced to the market.

Improved hardware and software capabilities are critical for supporting comfortable immersive user experiences that require spatial audio and high-definition (more than 4K per eye) video to create an accurate sensory perception of presence in a digital environment [1], [2], [3], [4], [5], [6]. Specifically, wireless immersive applications necessitate powerful and stable radio connectivity with very high throughput and ultra-low latency to the processing unit, and failure to satisfy these requirements leads to discomfort, disorientation, and nausea caused by human sensory conflict. Of course, these requirements can be already satisfied in test environments (i.e., with the assumption of very large bandwidth all dedicated to XR users) using state-of-the-art

technologies such as the latest third-generation partnership project (3GPP) 5G New Radio (NR) standard. However, with the growing penetration level and usage intensity of multi-user XR applications, available radio resources should be shared between the users, causing link qualities to deteriorate due to inter-user interference. Therefore, to meet future capacity demand, there is a need to further enhance the network capacity by introducing novel disruptive technology components.

One feasible solution for enabling the massive use of immersive applications is to exploit the abundant spectrum available in millimeter-wave (mmWave) bands, i.e., the Frequency Range 2 (FR2) in the 5G NR. The critical advantage of mmWave is the availability of wide bandwidth and the possibility of miniaturized antenna elements that enable multi-antenna systems with highly-directional data transmissions. However, with this promise of high throughput connectivity also comes with specific challenges, such as volatile channel quality, random blockage effect, and complex 3D interference footprint. A significant research effort is now carried out by academia and industry to address these critical issues, using, for example, multi-point connectivity across large antenna arrays [7].

Another possible solution for enabling multi-user wireless XR applications is using novel coded caching (CC) techniques that prioritize the content relevant to wireless bottleneck areas and enable the cumulative cache memory of users in the network to be used as an additional communication resource [8]. This solution is especially appealing as the onboard memory is becoming cheaper to implement and is available in larger quantities on modern devices. Caching is a well-studied concept; it has been used for a long time to place prevalent data closer to requesting users, reducing delivery time and network congestion probability [9], [10], [11], [12]. However, CC extends the benefits of traditional caching techniques by enabling a new performance gain proportional to the cumulative cache size in all users [8]. Interestingly, this new gain can also be combined with the multiplexing gain of multi-input single-output (MISO) [13], [14], [15], [16], [17], [18] and multi-input multi-output (MIMO) [19], [20], [21], [22], [23] communications using recently introduced multi-server and multi-antenna coded caching techniques.

Using caching techniques to improve users' quality of experience (QoE) in wireless XR applications is studied by a number of works in the literature. From a general perspective of caching, it is shown that utilizing the storage and computing capabilities of XR mobile gadgets could effectively alleviate the traffic burden over the wireless network. Moreover, significant bandwidth and delay-reduction gains have been demonstrated [10], [11], [24], [25]. Similarly, applying CC techniques in XR use cases has been considered by a handful of works [26], [27], [28], [29]. The general idea is to use two important features of wireless XR applications: 1) the content requested by users in such applications is used to reproduce their field of view (FoV) and hence is location-dependent, and 2) the size of the file

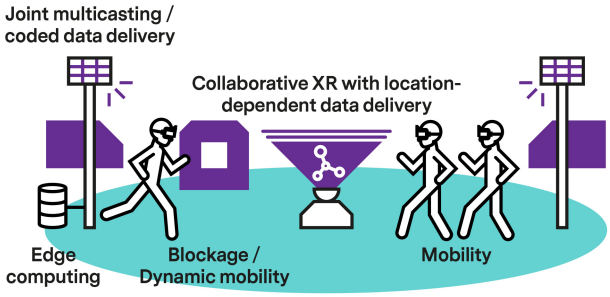

FIGURE 1. Immersive viewing scenario with coded caching.

library (i.e., the set of files that could be requested by users) is naturally limited. These features are employed to design new 'location-dependent' CC schemes, where non-uniform portions of the cache memory are allocated to store (parts of) the content files requested in different locations within the application hall, and novel CC techniques are employed to increase the achievable data rate for the resulting cache placement. The non-uniform cache allocation allows for storing larger portions of the content files requested in locations with poor connectivity in the cache; hence, users with bad channel conditions need to download smaller data amounts from the server and do not experience excessive data delivery delays. This results in improved QoE throughout the entire application environment by removing wireless connectivity bottlenecks. Due to their close alignment with the context of this paper, CC schemes of [26], [27], [28], [29] are reviewed in more detail in Section IV-D.

This paper takes a broader look, from both technical and standardization perspectives, at the benefits and problems of applying CC techniques, especially in the context of multi-antenna communications, to address critical communications bottlenecks of multi-user XR applications. In this regard, we first introduce our envisioned cache-aided multi-user XR setup and study recent multi-antenna CC advancements. Then, we review how standardization bodies, especially 3GPP, consider XR requirements in their specification studies and explore how our envisioned XR scenario relates to their foreseen use cases. Finally, we discuss in more detail the challenges arising while integrating CC techniques into multi-user XR applications and propose novel solutions to address the identified issues.

A. THE ENVISIONED XR SCENARIO

We envision an XR or hyper-reality environment where a large group of users is submerged into a network-based immersive application. The use case of the setup can be, for example, educational, industrial, gaming, defense, or social networking [30]. The XR application runs on high-end eyewear gadgets that require heavy multimedia traffic and are bound to guarantee a required level of user QoE within the operating theatre. A general illustration of such a scenario is shown in Figure 1.

Increasing computation capabilities of modern chipsets allow more and more processing to be handled locally

within high-end XR devices. However, with the stringent requirements of immersive applications and the form factor constraints of XR devices limiting the allowed heat dissipation level, full local processing of XR applications is still unattainable. This has led to the development of load-splitting models that enable major parts of the computation to take place in edge servers while offloading delay-sensitive final refinements to be handled locally at XR devices [30]. Such load-splitting models rely heavily on a reliable and fast communication channel between the edge and XR devices, as the servers should constantly be informed of the application environment and transfer their generated processing and rendering outputs to the end users [30]. In this regard, in our model, we assume that the transceivers on interfacing eyewear devices support high throughput demands, e.g., by operating on mmWave bands, and remain connected to the network, possibly through multiple transmit-receive points (TRPs) that enable reliable communication via multi-connectivity schemes to provide improved resilience to random blockage and increased coverage and capacity in the application area [31].

We also assume that multimedia consumers (i.e., network users) are scattered across the application area and can move freely, and their streamed data is unique and highly location- and time-dependent [26]. Like any other interactive scenario, our envisioned application also requires instant delivery of frequently changing data elements. This so-called *dynamic* data part includes, for example, the data related to the movements and actions of other users in a gaming application. However, a notably large part of the delivered content, e.g., the data required for rendering the background scenery of the user's FoV, is *static* (i.e., non-interactive) and can be cached beforehand when favorable radio channel conditions and excess communication resources are available (see Figure 2).[1] This feasibility of caching, together with the limited number of cachable files, provides the opportunity for efficient use of pooled memory resources through intelligent coded caching mechanisms [26], [27], [28]. The result is the possibility of delivering high-throughput, low-latency data traffic while providing high stability and reliability of radio connections for a truly immersive experience.

II. CODED CACHING (CC)

The pioneering work in [8] introduced a novel *coded caching* scheme where instead of merely replicating high-popularity content near (or at) end-users, fragments of the contents were spread across different cache memories throughout the network. CC works in two phases, *placement* and *delivery*. During the placement phase, content files are split into smaller parts, and these parts are stored in the cache memories of different users. This phase is performed, for example, when the network traffic is low (to avoid congestion) or when the users are close to TRPs (to minimize the transmission time and energy expenditure). Then, during the delivery phase, after the network users reveal their requests, several codewords are built and each codeword is multicast to a subset of target users.

A simple illustration of a CC-aided operation with a single-antenna transmitter is shown on the left-hand side of Figure 3. In this simplified scenario, two users, each with cache memory large enough to store one of the files A and B, request files from a single-antenna server over a shared link. There is no prior information about the request probabilities. It can be shown that with a classic caching solution (i.e., no coded placement or delivery), the worst-case load on the shared link cannot be smaller than the size of one file. However, using CC, we can halve this worst-case load by allowing each transmission to serve two users

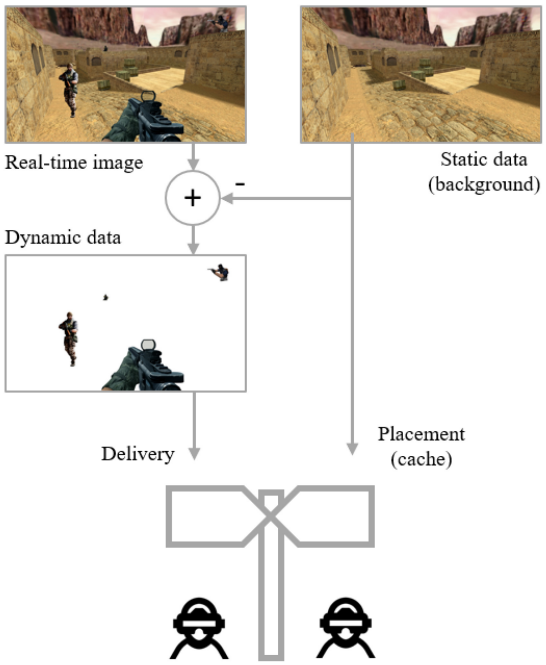

FIGURE 2. Data decomposition into static and dynamic parts (screenshots are from the counter-strike game).

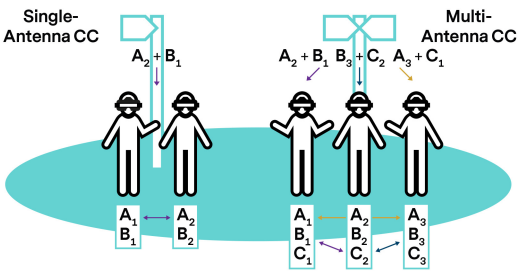

FIGURE 3. Single- and multi-antenna coded caching.

1. In typical virtual gaming applications, even the dynamic part of the FoV can be described by smaller, cacheable elements. For example, a moving person in an XR game can be represented as a superposition of multiple well-defined geometrical shapes covered with various textures and possibly overlaid by the avatar of the corresponding player. All these elements (geometrical shapes, textures, and avatars) are cacheable and can be efficiently delivered using coded caching techniques (i.e., by storing part of the elements and multicasting the rest). Of course, one would still need to transmit control/instruction data describing how to reconstruct the dynamic part from both the cached elements and the multicast data. More details can be found in [1], [4].

TABLE 1. Request patterns, transmitted data, and link loads in classic and coded caching, for the example network.

		Classic Caching				Coded Caching			
requests		cache contents		transmitted data	link load	cache contents		transmitted data	link load
user 1	user 2	user 1	user 2			user 1	user 2		
A	B	A	B	-	0	A_1, B_1	A_2, B_2	$A_2 \oplus B_1$	0.5
B	A			B, A	2			$B_2 \oplus A_1$	0.5
A	A			A	1			$A_2 \oplus A_1$	0.5
B	B			B	1			$B_2 \oplus B_1$	0.5
Average Link Load:					**1**	**Average Link Load:**			**0.5**

simultaneously. For the example network, this is achieved by splitting each file into two equal-sized parts, and caching A_1 and B_1 at user one and A_2 and B_2 at user two. Then, for example, if users one and two request files A and B, respectively, we simply transmit $A_2 \oplus B_1$, where $\oplus$ denotes the XOR operation over the finite field. This way, users one and two can remove undesired terms B_1 and A_2 from the received signal using their cached contents and decode A_2 and B_1, respectively. For better clarification, in Table 1, we have compared both classic and coded caching schemes for this example network, considering all possible request patterns. In general, using CC, the number of users served with each transmission (hence, the reduction in the worst-case load) scales with the cumulative cache size in the network [8]. In mathematical terms, if each user can cache a portion γ of the entire library and there exist K users in the network, we can serve $t+1$ users in each transmission, where $t = K\gamma$ is called the *coded caching gain*.

Following the original CC scheme in [8], many ensuing works in the literature applied its core idea to more diverse network setups. An important direction was to apply CC techniques in MISO setups, revealing the exciting result that the CC and spatial multiplexing gains are additive. In fact, with coded caching gain t, if the spatial multiplexing gain of L is attainable by the transmitter,[2] the total number of users served in each transmission can reach $t+L$. In [17], it is shown that this number is optimal under the conditions of uncoded cache placement and single-shot data delivery (i.e., when the decoding process of one transmission does not depend on other transmissions). As a clarifying example, consider the MISO setup in the right-hand side of Figure 3, where three single-antenna users, each with a cache memory large enough to store one file, request data from a multi-antenna server that can attain a spatial multiplexing gain of $L = 2$. Requests are made from a library of three files A, B, and C (i.e., $\gamma = 1/3$ and $t = 1$), and there is no prior knowledge of the requests. For this network, by splitting each file into three equal-sized parts and caching them as shown in Figure 3, we can serve all $t+L = 3$ users simultaneously (while without CC, at most $L = 2$ users can be served). To achieve this additional gain, we create an appropriate XOR codeword for every group of users of size two, and design multicast beamformers to suppress the interference caused by each of these group-specific codewords at the other user not belonging to that group. For example, assuming users 1-3 have requested files A, B, and C, respectively, the transmission vector for this network is built as

$$\mathbf{x} = (A_2 \oplus B_1)\mathbf{v}_3 + (A_3 \oplus C_1)\mathbf{v}_2 + (B_3 \oplus C_2)\mathbf{v}_1, \quad (1)$$

where $\mathbf{v}_k$ is the beamforming vector designed to suppress interference at user k.[3] Let us review the decoding process at user 1, after the transmission of $\mathbf{x}$. Denoting the channel vector of this user and the additive noise at its receiver with $\mathbf{h}_1$ and z_1, respectively, we can model the received signal as $y_1 = \mathbf{h}_1^T\mathbf{x} + z_1$, i.e.,

$$y_1 = (A_2 \oplus B_1)\mathbf{h}_1^T\mathbf{v}_3 + (A_3 \oplus C_1)\mathbf{h}_1^T\mathbf{v}_2 + \underline{(B_3 \oplus C_2)\mathbf{h}_1^T\mathbf{v}_1} + z_1. \quad (2)$$

According to the definition, the interference caused by the underlined term in (2) is suppressed by beamforming vector $\mathbf{v}_1$. Hence, user 1 can decode the two codewords $(A_2 \oplus B_1)$ and $(A_3 \oplus C_1)$ from y_1, using, e.g., a successive interference cancellation (SIC) receiver. Finally, as user 1 has B_1 and C_1 in its cache (see Figure 3), it can extract its desired terms A_2 and A_3 out of the codewords readily using an XOR operation.

A. CODED CACHING WITH SIGNAL-LEVEL INTERFERENCE CANCELLATION

In general, CC techniques rely on spreading content fragments over the cache memories throughout the network. These small fragments are usually called *subpackets* in the literature, and the process of splitting content files into subpackets is known as *subpacketization* [35]. For example, for the single- and multi-antenna setups in Figure 3, content files are split into two and three subpackets, respectively. The problem is that, for both baseline single-antenna [8] and multi-antenna [14] CC schemes, the number of subpackets

2. The spatial multiplexing gain refers to the total number of parallel streams (across multiple users) that can be handled by a single access point, and its value is upper-bounded by the number of antennas at that access point. While with the recent emergence of communications in higher frequencies (e.g., in mmWave bands), it has become possible to employ larger antenna arrays (> 100) at both transmitters and receivers, the attainable multiplexing gain is more constrained due to practical limitations in the number of RF chains and available baseband processing and transmit power, as well as severely constrained pilot resources for channel sounding reference symbols [32], [33], [34]. Of course, as shown in [15], MISO CC techniques can significantly benefit from improved multicast beamforming gains when the spatial multiplexing gain is smaller than the antenna count.

3. As in conventional MISO systems without coded caching, various strategies can be used to design beamformers, and they affect the system performance depending on the operating signal-to-noise ratio (SNR). A detailed discussion on beamforming for MISO coded caching schemes is provided in [15].

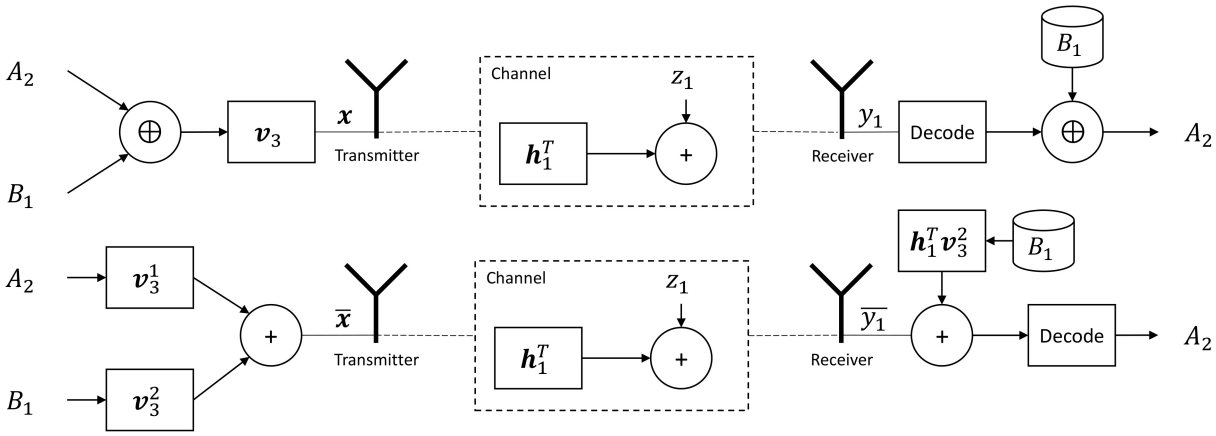

FIGURE 4. Bit-level (top) and signal-level (below) schemes.

(i.e., the subpacketization value) grows exponentially with the number of users, severely limiting the real achievable CC gain in practice [35].

While single-antenna setups are not flexible for reducing the subpacketization value [36], multi-antenna setups can work with modest subpacketization values, even smaller than their comparable single-antenna counterparts [35]. This is enabled by a new cache-aided interference cancellation mechanism, where unwanted terms are regenerated from the local memory and removed *before* the received signal is decoded at the receiver. We use the term *signal-level* approach for this new mechanism and denote the classic cache-aided interference cancellation mechanism (after decoding at the receiver) as the *bit-level* approach.

For clarification, let us review how the signal-level approach could be applied to the MISO setup in Figure 3. For this network, instead of transmitting $\mathbf{x}$ in (1), one can use

$$\bar{\mathbf{x}} = A_2\mathbf{v}_3^1 + B_1\mathbf{v}_3^2 + A_3\mathbf{v}_2^1 + C_1\mathbf{v}_2^2 + B_3\mathbf{v}_1^1 + C_2\mathbf{v}_1^2 \qquad (3)$$

to deliver the same data terms to every user. In (3), $\mathbf{v}_k^1$ and $\mathbf{v}_k^2$ denote two (possibly different) beamformers, both suppressing data at user k. For clarification, let us review the decoding process at user 1. Following the same discussions as before, this user receives

$$\begin{aligned}\bar{y}_1 = {} & A_2\mathbf{h}_1^T\mathbf{v}_3^1 + B_1\mathbf{h}_1^T\mathbf{v}_3^2 + A_3\mathbf{h}_1^T\mathbf{v}_2^1 \\ & + C_1\mathbf{h}_1^T\mathbf{v}_2^2 + \underline{B_3\mathbf{h}_1^T\mathbf{v}_1^1} + \underline{C_2\mathbf{h}_1^T\mathbf{v}_1^2} + z_1, \qquad (4)\end{aligned}$$

where the interference from underlined terms is suppressed by beamformer vectors $\mathbf{v}_1^1$ and $\mathbf{v}_1^2$. Now, user 1 has to first regenerate the remaining interference terms $B_1\mathbf{h}_1^T\mathbf{v}_3^2$ and $C_1\mathbf{h}_1^T\mathbf{v}_2^2$ using its cache contents[4] and remove them in the signal domain from the received signal before it can decode its desired terms A_2 and A_3. This contrasts the bit-level approach where the cache contents were used after the received signal was decoded at the receiver. A graphical comparison of bit-level and signal-level approaches for the considered example network is provided in Figure 4. Note that in this figure, only the encoding process of A_2 and B_1 and the decoding process of A_2 at user 1 are shown.

Although the signal-level approach incurs a noticeable performance loss at the finite-SNR regime due to its inferior multicasting gain compared with the bit-level approach [37], [38], it provides great flexibility in addressing many practical bottlenecks of CC schemes, and as a result, is thoroughly studied in the literature. For the sake of brevity, we briefly review a few notable research directions here:

➢ With signal-level interference cancellation, it is possible to reduce the subpacketization requirement while serving the same number of $t + L$ users as bit-level schemes in each transmission [35]. It is even shown that the exponential subpacketization growth (with respect to the network size) in bit-level schemes can be replaced by linear scaling in MISO networks with a large spatial multiplexing gain at the transmitter [39].

➢ With signal-level interference cancellation, the complex multi-group multicast beamformer design of bit-level schemes [15] could be replaced with a much simpler multi-user unicast beamformer design [39], [40]. This allows us to design CC schemes that are applicable to very large networks and perform well in the finite-SNR regime.

➢ Signal-level interference cancellation allows designing CC schemes for dynamic networks where the users can join and leave the network freely [40], [41], [42]. Such dynamic schemes are based on the shared-cache model [43], [44], and are described in more detail in Section IV-E.

➢ In [21], a signal-level CC scheme is introduced for MIMO setups (with multiple antennas at both the transmitter and receivers) that provides a multiplicative boost in the CC gain compared to MISO setups with a very low subpacketization overhead.

We should note that in addition to performance and flexibility, bit- and signal-level CC schemes are also different in signaling requirements and the way they affect the physical layer. This is elaborated on in Section IV-G.

4. The effective channel coefficients $\mathbf{h}_1^T\mathbf{v}_3^2$ and $\mathbf{h}_1^T\mathbf{v}_2^2$ can be estimated, for example, from demodulation reference signal (DMRS) pilots.

III. EXTENDED REALITY AND STANDARDIZATION

XR is expected to be one of the key 6G drivers [7]. Various industrial, educational, gaming, and social networking wireless XR applications will emerge in the coming years, and they will push the capabilities of the underlying communication infrastructure. Accordingly, XR-related technical discussions are already in progress in various standardization bodies such as 3GPP. To date, classifications and requirements of XR applications and their interconnection with 3GPP terms and standards are studied, and the results are made available in various technical reports (TR). More precisely, XR and cloud gaming (CG) evaluation methodology and performance were studied in 3GPP Release 16, with TR 38.838 [45] providing comprehensive performance analysis for XR and CG in relevant 5G NR deployment scenarios. This work continued in Release 17 within new study items with a focus on improvements for XR. The results of these study items are reported, e.g., in TR 38.835 [46]. Finally, in Release 18, NR enhancements for facilitating XR are standardized. This includes, e.g., radio access networks XR awareness operations, and XR-specific enhancements for power saving and capacity [47].

In this section, we briefly overview the comprehensive report in TR 26.928 [30], which provides a general introduction to XR terms and definitions, classifies XR device form factors and use cases, and clarifies the technical requirements of various XR applications. Then, in Section IV, we specifically address practical implementation and specification challenges for enabling CC techniques in multi-user XR environments.

A. XR FORM FACTORS AND USE CASES

XR device classification is done according to the physical form (flat screen, mounted headset, eyewear), processing unit location (external or inbuilt), and cellular communication unit location (external or inbuilt). These properties affect the amount of power that can be safely dissipated and hence, the achievable communication and processing power in the device. A schematic description of the XR device classification, as categorized in [30], is shown in Figure 5. In this classification, depending on the application type, which can be either augmented reality (AR) or virtual reality (VR), XR devices are put into two broad categories. In both categories, it is possible to use the smartphone screen as the main display (denoted by XR5G-P1 in Figure 5). Otherwise, if a head mount device (HMD) is used as the display to enhance the user experience, the categorization is based on: 1) where the required XR processing at the user side is performed and 2) where the 5G transceiver is located. Both processing and 5G connectivity can be located inside or outside the HMD hardware, as shown in the figure. In the case of AR, the difference between XR5G-A3, XR5G-A4, and XR5G-A5 is the device form factor: as more advanced technology becomes available, we expect to move from a bulky HMD (XR5G-A3) to a small device with minor differences to normal eyeglasses (XR5G-A5). More detailed explanations can be found in [30].

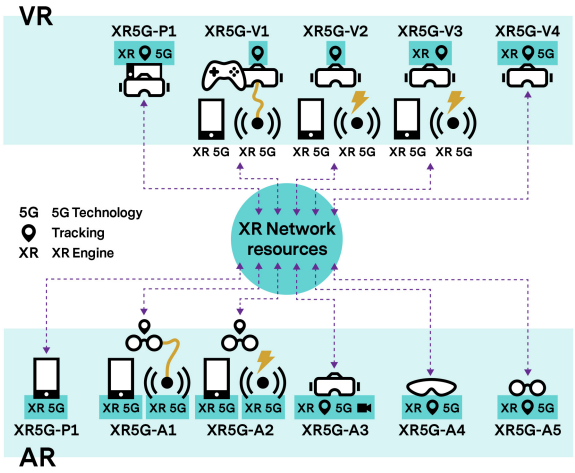

FIGURE 5. Device form factors as categorized in [30].

There also exists a thorough discussion in [30] on possible XR use cases and how future networks can help realize them. These use cases lie in seven broad categories, as shown in Table 2. Our considered scenario, as explained in Section I-A, can cover various use cases from different categories. One prominent example is an XR gaming application, where the users wear XR headsets and are physically co-located in the same application environment. We will discuss the use cases that can benefit from CC in more detail in Section IV.

B. XR QUALITY OF SERVICE (QOS) REQUIREMENTS

XR applications should provide the user with the feelings of *immersion* and *presence*, defined as the feelings of being surrounded by, and physically and spatially located in the virtual environment, respectively. Providing such feelings necessitates very fast and accurate position and orientation tracking and very high-data-rate and low-latency communications. Putting in numbers, an acceptable XR experience requires sub-centimeter positioning accuracy, quarter-degree rotation tracking accuracy, 8K per-eye video resolution, and motion-to-photon latency in the order of 50 milliseconds (the exact number varies by the application category) [30]. Of course, the 5G standard already includes a QoS provision model using the 5G QoS identifier (5QI) parameter. However, many open questions still exist on how the general and graphical processing operations should be split among the client devices and edge servers and which 5QI parameters should be used for each application category.

C. RENDERING AND PROCESS SPLIT

In the context of XR, we use the term rendering to refer to the process of generating animated 3D graphics from abstract computer-language models. The software responsible for rendering is called the rendering engine and is part of the higher-level XR engine, which is the software

TABLE 2. XR use case categories and their examples, as provided in [30].

Use case category	Use case example
Offline sharing of 3D objects	Alice downloads a 3D model of a sofa from an online shop. She puts the virtual representation of the sofa in her house and checks if it fits nicely.
Real-time XR sharing	Alice downloads a 3D model of a sofa and puts the virtual representation of the sofa in her house. Then, she initiates a live discussion with a friend about that.
XR multimedia streaming	Bob is watching a live sports game while wearing an XR headset. He can choose where he sits in the stadium and interact with the people sitting nearby.
Online XR gaming	Alice and her friends engage in a multiplayer XR game. Each of them can be physically present in a dedicated application hall or just engage remotely.
XR mission critical	Police has an important mission. Officers engaging in the mission get live data overlaid on their AR glasses. The data depends on their location and helps with the mission.
XR conferencing	A scientific conference is held in a hybrid physical/virtual mode. All the participants wear XR headsets and have the perception of really being at the same place.
Spatial audio multiparty call	Bob initiates a group call with his friends. Each participant can share his/her spatial audio and others hear it as if they were physically there.

development environment used for building XR applications. The two best-known rendering techniques are rasterization and ray-tracing, whose technical explanation is out of scope for this paper. Ray-tracing generates more realistic outputs but is also more computationally intensive. The selection of the rendering technique and various parameters that can be tuned for the selected technique affects the type and number of data buffers to be processed and the order in which they should be processed. As a result, it also affects the possibility of splitting the rendering process among the local XR device and the edge infrastructure, which is vital for XR applications as XR devices generally have limited processing capability due to limitations in processing power, battery capacity, and allowed power dissipation level.

An important question in process splitting, affecting both the energy consumption and user QoE, is where to generate the final image of the FoV shown to the user's eyes. Fully rendering at end devices is practically infeasible due to the large energy expenditure of the underlying processing, and offloading all the rendering to the edge server makes the QoE prone to the smallest variations in communication latency. Solutions to this issue include splitting the rendering process between the edge and the device or rendering mainly at the edge and making final corrections (following the user's instantaneous pose) at the end device [30]. In this paper, we promote a modified version of the latter solution: we envision using the edge server for rendering an omnidirectional 3D representation of the static part of the content. The generated 3D images are then transferred, using efficient multi-antenna coded caching techniques, to the end devices which are responsible for overlaying the dynamic and static parts as well as making the final pose corrections locally (see Figure 2).[5] Of course, such a process splitting architecture necessitates a higher energy consumption for processing than fully rendering at the edge, as well as larger communication overheads for delivering the 3D content [30]. Nevertheless, the QoE would be more robust to latency variations compared with full rendering at the edge (as minor instantaneous corrections are carried out locally), and the energy expenditure of the underlying processing would be much smaller than full rendering at the device. Moreover, the communication overhead could be efficiently alleviated using novel multi-antenna CC techniques.

5. As explained earlier, a notable part of the dynamic content can also be delivered with coded caching techniques. Also, overlaying dynamic and static parts can be done, e.g., as explained in [1].

IV. CODED CACHING FOR ENHANCED XR

Ideally, multi-antenna CC has great potential to provide significant performance improvements for next-generation XR applications. However, critical practical impediments should be resolved before it can be deployed in practice and/or considered for standardization. The most prominent issue is the lack of a proper framework defining the content structure and the timeline for placement/delivery phase operations. In this section, we first review a few application use cases that lie within our envisioned system model in Section I-A. Then, we propose a possible framework for integrating CC techniques in such use cases and review a few issues open for future study.

A. CC AND APPLICATION USE CASES

CC suits well to XR applications where the users have physical proximity, can move freely, and require location-dependent data. Moreover, the requested content (or part of it) should be cacheable by nature (as thoroughly discussed in Section I-A). Some examples of XR applications where CC can be applied are:

➢ **An XR gaming application** where the users are scattered in the game hall, equipped with XR headsets, and move freely. The users play an interactive game, and their requested data at any moment depends on their location at that moment;

➢ **An XR museum application** where the users entering the museum are given XR headsets and move freely throughout the museum building. Close to each historical

TABLE 3. Theoretical CC gains in different XR setups, 50*cm* × 50*cm* STU size, 100 MB of 3D image size.

Parameter	Scenario I	Scenario II	Scenario III	Scenario IV	Scenario V
Application environment size	$5m \times 5m$	$5m \times 5m$	$5m \times 5m$	$10m \times 10m$	$10m \times 10m$
User count	5	10	10	10	40
Library file count	100	100	100	400	400
User cache size	4 GB	4 GB	8 GB	8 GB	8 GB
The coded caching gain (t)	2	4	8	2	8
CC packet size	5 MB	10 MB	20 MB	20 MB	$\sim$ 20 KB
Transmitter spatial multiplexing gain (L)	2	2	2	2	2
Parallel streams w/ CC ($t + L$)	4	6	10	4	10
Improvement by CC	**100%**	**200%**	**400%**	**100%**	**400%**

item, nearby users may jointly enter a related virtual 3D world where they can interact with virtual elements;

➢ **An XR conferencing application** where the users may join physically or virtually. In both cases, the users are equipped with XR headsets and move freely among different halls. The users can interact with each other and with the environment.

All these use cases share common properties; the users have physical proximity, they can move freely and require location-dependent data, and the interactive nature of the application necessitates ultra-low communication latency. However, even with recent advancements, it is challenging to simultaneously transmit very high data rate 3D streams with ultra-low latency to multiple users. The situation becomes even more challenging as the users start to move, as the stringent requirements on rate and latency should be met throughout the whole application environment.

B. RENDER AND PROCESS SPLIT FOR CC-AIDED XR

In Section III-C, after briefly reviewing render and process split frameworks considered by 3GPP, we promoted a model where rendering the omnidirectional 3D representation of the static part of the content is done at the edge server, and overlaying the dynamic part and final pose corrections are done locally at the end device. This model makes the QoE less prone to small variations in transmission delay but requires delivering larger amounts of data, which could be alleviated using multi-antenna CC techniques.

We must emphasize that the promoted render and process split model is *not* the only possible way to use CC techniques to enhance QoE in XR applications. In fact, we may always use CC techniques when (part of) the delivered content is cacheable in nature. Indeed, there exist multiple works in the literature for efficient rendering and delivery of omnidirectional image and video for XR use cases, with a comprehensive survey provided in [48]. For example, considering the tile-based rendering model where an omnidirectional frame is split into multiple tiles and only tiles relevant to the FoV of the user are delivered [48], the static background of the image in each tile is again cacheable and could be efficiently delivered with CC techniques. Even if we intend to deliver the tiles closer to the center of the FoV with a higher quality to save bandwidth (as those tiles have a more prominent effect on the QoE of the user), we may use novel CC techniques, e.g., based on multiple descriptor codes (MDC) [49], [50], to support different image qualities for various tiles.

Similarly, as long as (part of) the transmitted data is cacheable in nature (or could be split into cacheable elements as discussed for the dynamic part of the content in Section I-A), CC schemes provide benefits even if disruptive techniques such as semantic communication are used to deliver the content [51], [52], or if multi-sensory XR techniques are applied to further improve users' immersion into the virtual world. For example, with semantic communications, the cacheable part could include the semantic language database.

C. THE ENVISIONED XR FRAMEWORK

As discussed earlier, for the considered collaborative XR setup, we propose and using multi-antenna CC techniques to deliver them to the end users efficiently. As discussed, these cacheable parts include, for example, the static part of the FoV as well as the building elements of the dynamic part. To keep discussions simple, here we assume CC-aided data delivery is applied only to the static part. To reduce the impact of transmission delay variations on QoE, we assume omnidirectional 3D images are used to represent the static part of the content, and final pose corrections are done at the end device. In this regard, we define the **single transmission unit (STU)** to be the smallest area unit for which the static view can be transmitted with a single 3D image. For example, we can assume the XR application environment is split into tiles of size 50×50 cm and each tile is an STU, i.e., the user needs to receive a new 3D image from the server to reconstruct the background scenery as it moves from one tile to another (see Figure 6).

Delivering high-definition omnidirectional 3D images in a collaborative XR application necessitates transmitting large amounts of data (hundreds of megabytes) in a very short time frame (tens of milliseconds). Nevertheless, this burden could be handled through the efficient usage of novel multi-antenna CC techniques. Table 3 represents theoretical gains (i.e., upper bounds on the achievable performance) in terms of the number of possible interference-free parallel data streams resulting from CC techniques in a few example

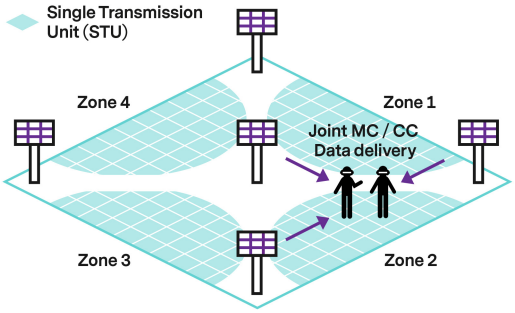

FIGURE 6. Zoning and STUs for improved XR.

XR setups. This table highlights why CC could be an exciting solution for future XR setups: it enables a large performance gain under realistic assumptions, and this gain scales with the number of users. This is in contrast to many other communication techniques, where the performance primarily deteriorates as the network size scales.[6]

One might argue the feasibility of the proposed framework in very large XR environments (e.g., within a large exhibition or conference hall) as, in such a case, users could be far from each other and served by different transmitters. For this case, we further propose splitting the whole environment into several **zones**, where the users within the same zone could participate in a CC-aided content delivery session and the cache contents of users are updated as they move between the zones (see Figure 6). Of course, one should note that although zoning allows allocating larger cache portions to each cached file (as it limits the number of files by constraining the physical area), it does not necessarily increase the achievable CC gain. This is because the CC gain is proportional to the total cache size of the users in the same CC-aided delivery session [8], and zoning limits the expected number of such users.

Another important issue is the overall *energy efficiency* of the proposed framework. Of course, studying the energy efficiency of the CC schemes is not limited to XR setups and is addressed in a limited number of works in the literature. One notable example is [53], where the energy efficiency of the delivery phase is compared with and without using CC schemes. However, for our proposed scheme, studying energy efficiency should also consider the placement phase, as the users have to download large amounts of data with corresponding energy expenditure to fill up their cache memories, while part of the downloaded content might never be used in the delivery phase. The zoning further increases the placement cost, as the users might need to update their cache contents frequently while moving between the zones. One way to address this issue is to optimize the zoning process such that the users pass through the vicinity of TRPs as they move from one zone to another; so that they can promptly download the data required for updating their cache contents at high rates and hence causing limited overall energy expenditure. Moreover, it is also possible to optimize the XR application (using the available information on how the zoning is performed) to increase the expected time each user resides within a single zone, thus decreasing the need to update the cache contents and increasing the number of time cache contents are reused. As a clarifying example, consider an XR application where two groups of players are engaged in a first-person action game. Then, noting that the users spend more time in designated conflict areas in such an application, the game map could be altered such that each conflict area lies within a single CC zone.

6. It should be noted that the numbers in Table 3 are indicative; used only to provide a general idea of the gains possible by CC techniques. Also, to calculate the CC packet size, we have considered the CC scheme of [35] for its reduced subpacketization (except for Scenario I, where the CC scheme of [14] is used as the scheme of [35] incurs a performance loss).

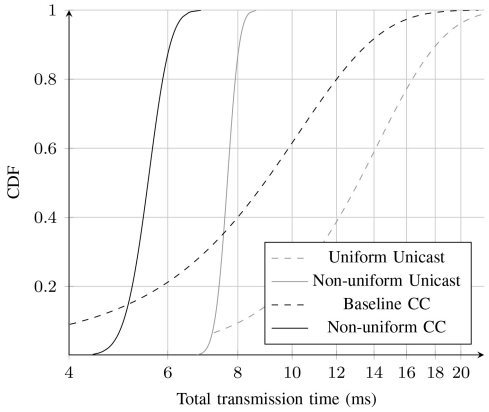

FIGURE 7. The CDF of total delivery time (logarithmic-scale) for $K = 36$ users, $\gamma = 0.33$, and $L = 6$. The variance of the shadowing effect of the channel is $\sigma_s = 7$ [29].

As a final note, although most CC schemes in the literature assume synchronized user requests (which is not a strong assumption for XR use cases), it is also possible to apply CC schemes without such a constraint. For example, one may use the CC scheme in [54], for which the achievable performance is within a multiplicative factor of two from the case of the synchronized request under the assumption of uncoded cache placement.

D. LOCATION-DEPENDENT CODED CACHING FOR XR

As discussed in Section I, XR applications possess two important features: location-dependent content requests and file libraries of limited size. These features are used in [26], [27], [28], [29] to design new CC schemes well-tailored to XR setups. The core idea is to allocate larger cache portions for storing (parts of) the content requested in STUs with poor channel connectivity to avoid excessive content delivery delays and improve QoE. Then, novel CC techniques are introduced to provide a global caching gain for the resulting non-uniform memory allocation. Of course, multi-user content delivery in such schemes requires delivering different-sized data chunks to various users within a single transmission, which is done using either nested code modulation (NCM) [55] or a high-performance beamformer design [29]. Here, to showcase the QoE improvements brought to XR applications by CC techniques, in Figure 7,

we have provided simulation results from [29] to compare the following four schemes: 1) the location-dependent CC scheme of [29], denoted by 'Non-uniform CC,' 2) the baseline multi-antenna scheme of [14], denoted by 'Baseline CC,' 3) a reference scheme with non-uniform memory allocation but unicast content delivery (i.e., no CC technique used), denoted by 'Non-uniform Unicast,' and 4) another reference scheme with uniform memory allocation and unicast content delivery, denoted by 'Uniform Unicast.' This figure clarifies the general impact of coed caching techniques as well as the QoE benefits of location-dependent CC schemes: the variance in the content delivery time is much smaller than the baseline CC scheme due to the underlying non-uniform memory allocation, and the achievable rate is better than the unicast case due to the global caching gain of CC techniques.

Of course, we should note that other works also exist in the literature on location-dependent coded caching [56]. However, in this paper, we have limited our review to [26], [27], [28], [29] as they also consider XR applications as the use case.

E. CODED CACHING FOR DYNAMIC SETUPS

From the cache placement design perspective, CC schemes range from fully centralized schemes, where a central server instructs what should be cached by every individual user, to fully decentralized ones, where the users randomly cache arbitrary portions of the content files. Fully decentralized schemes have the important advantage of being flexible to variations in network parameters but provide comparable performance to fully centralized schemes only asymptotically (i.e., when the number of users is very large) [57]. This has led to the design of *shared-cache* CC models, which are centralized in the sense that the cache contents of the users are determined by a set of predefined *cache profiles* but decentralized in the sense that the assignment of the users to cache profiles could be random (it is even possible that multiple users are assigned to the same profile and have similar cache contents) [43], [44], [58], [59], [60]. All shared-cache CC models are signal-level CC schemes and have similar strong (e.g., reduced subpacketization) and weak (e.g., inferior finite-SNR performance) points. Due to their great flexibility, shared-cache models have also been used to build CC schemes for dynamic networks where the users may join and leave the network freely, with thorough analyses available in the literature for both high-SNR [41], [42] and finite-SNR [40] regimes. In Figure 8, to showcase the performance of dynamic CC schemes, we have provided simulation results from [41] that compare the performance of their proposed scheme with the two baseline schemes of unicasting only (no CC scheme applied, denoted by Unicast Only) and no dynamic conditions (CC scheme of [40] applied, denoted by Uniform CC). As can be seen, the performance of the dynamic CC scheme lies within the two baselines and gets better as users are assigned to the cache profiles more uniformly.

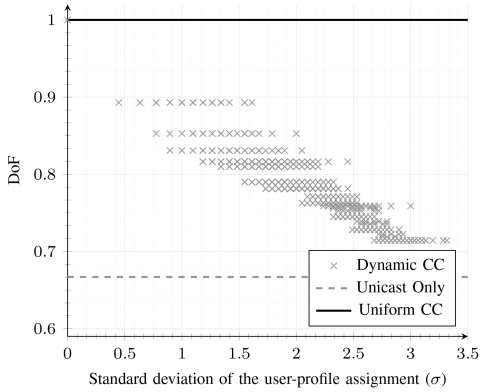

FIGURE 8. The average number of parallel streams per transmission (DoF) vs non-uniformity in the user-profile assignment. $K = 50$ users, $\gamma = 0.1$, $L = 9$, ten cache profiles.

In the context of CC schemes for wireless XR applications, even if the number of users is assumed to be fixed and known during the application runtime, we still face a dynamic setup for two good reasons: 1) with the zoning definition, CC data delivery is performed within each zone separately, and the number of users within a zone is not fixed as the users move between the zones, and 2) it is possible that a subset of users do *not* initiate any data requests in a time interval, e.g., if they have not changed their location from the previous interval. As a result, it is worthwhile to design dynamic CC schemes specially tailored to XR setups.

F. FURTHER PERFORMANCE IMPROVEMENTS

D2D-assisted coded caching: The CC gain is a result of cache-aided interference cancellation at the receivers. However, from another viewpoint, this interference removal can be regarded as *exchanging* data between multiple users. For example, in the single-antenna CC example in the left-hand side of Figure 3, it can be imagined that the transmitted codeword $A_2 \oplus B_1$ has enabled user one to deliver B_1 (cached in its memory) to user two, and at the same time, user two to deliver A_2 to user one. Now, instead of using multicasting to deliver the codeword $A_2 \oplus B_1$, two nearby users can exchange the requested terms using a direct device-to-device (D2D) link with higher capacity and lower latency [61], [62], [63].

Optimizing the zoning process: As discussed, zoning is a necessary tool for applying the proposed framework to large XR environments, and the way we perform zoning affects the energy efficiency of the system. In this regard, optimizing the zoning process to ensure that the users pass through the vicinity of TRPs while changing zones could improve the energy efficiency of the system noticeably. An important question arising here is whether to allow any overlap among the zones. With overlaps, the number of cache update operations might be increased, but users have more time to update their cache contents as they move between the zones, enabling more efficient resource allocation for the content update process and reducing the probability of incomplete cache updates. Studying the trade-off between the performance and energy efficiency of the network as the

size of the overlap areas is increased is out of the scope of this paper. Also, as mentioned in Section IV-C, for a given zoning, we may also improve energy efficiency by optimizing the XR application such that the users are less likely to change their zones within short time intervals. The result is an improved reuse factor of the cache contents, as the users spend more time and request a larger number of contents within each zone. Indeed, such optimizations of the XR application also affect the above-mentioned trade-off.

G. SELECTING THE PROPER SCHEME

An important implementation aspect is the choice of a suitable CC scheme for the target XR application scenario, as the schemes vary widely from the performance, complexity, and signaling overhead perspectives. Especially the choice determines the underlying interference cancellation approach. Unlike signal-level schemes, bit-level schemes allow the cache-aided interference cancellation to be carried out at higher network layers, making the process transparent to the physical layer and hence, easier to implement. However, as discussed earlier, the bit-level approach is more sensitive to the subpacketization problem, leading to either reduced CC gain or an extreme number of subpackets with increased complexity on signaling and the total number of individual transmissions.

On the other hand, signal-level schemes highly impact the physical later as the interference cancellation is performed before decoding the data, requiring the receiver to fetch the data to be canceled from its memory and to regenerate a replica of the expected interference by carrying out channel encoding, rate matching, scrambling, and modulation locally. The receiver also needs to estimate the effective channel, incorporate the effect of the receiver beamformer and equalizer, and convolute the generated signal with the effective channel for all interferers to be canceled. Nevertheless, the signal-level approach can facilitate a user-specific link adaptation in terms of modulation and coding and hence, may provide a way to alleviate the near-far problem hindering the CC operation.

V. CONCLUSION

We explored novel multi-antenna coded caching techniques as an appropriate candidate for resolving wireless connectivity challenges of future collaborative XR applications. We first reviewed recent advancements in multi-antenna coded caching techniques and then discussed how XR application requirements are addressed within the 3GPP framework. Finally, we identified new challenges arising from integrating CC techniques into multi-user XR scenarios and proposed novel solutions to address them in practice.

REFERENCES

[1] E. Thomas, E. Potetsianakis, T. Stockhammer, I. Bouazizi, and M.-L. Champel, "MPEG media enablers for richer XR experiences," 2020, *arXiv:2010.04645*.

[2] T. Taleb et al., "Toward supporting XR services: Architecture and enablers," *IEEE Internet Things J.*, vol. 10, no. 4, pp. 3567–3586, Feb. 2023.

[3] C. Chaccour, M. N. Soorki, W. Saad, M. Bennis, and P. Popovski, "Can terahertz provide high-rate reliable low-latency communications for wireless VR?" *IEEE Internet Things J.*, vol. 9, no. 12, pp. 9712–9729, Jun. 2022.

[4] K. Boos, D. Chu, and E. Cuervo, "Demo: FlashBack: Immersive virtual reality on mobile devices via rendering memoization," in *Proc. 14th Annu. Int. Conf. Mobile Syst. Appl. Serv.*, 2016, p. 94.

[5] M. Chen, W. Saad, and C. Yin, "Virtual reality over wireless networks: Quality-of-service model and learning-based resource management," *IEEE Trans. Commun.*, vol. 66, no. 11, pp. 5621–5635, Nov. 2018.

[6] E. Bastug, M. Bennis, M. Medard, and M. Debbah, "Toward interconnected virtual reality: Opportunities, challenges, and enablers," *IEEE Commun. Mag.*, vol. 55, no. 6, pp. 110–117, Jun. 2017.

[7] N. Rajatheva et al., "White paper on broadband connectivity in 6G," 2020, *arXiv:2004.14247*.

[8] M. A. Maddah-Ali and U. Niesen, "Fundamental limits of caching," *IEEE Trans. Inf. Theory*, vol. 60, no. 5, pp. 2856–2867, May 2014.

[9] G. S. Paschos, G. Iosifidis, M. Tao, D. Towsley, and G. Caire, "The role of caching in future communication systems and networks," *IEEE J. Sel. Areas Commun.*, vol. 36, no. 6, pp. 1111–1125, Jun. 2018.

[10] Y. Sun, Z. Chen, M. Tao, and H. Liu, "Communications, caching, and computing for mobile virtual reality: Modeling and tradeoff," *IEEE Trans. Commun.*, vol. 67, no. 11, pp. 7573–7586, Nov. 2019.

[11] Y. Sun, Z. Chen, M. Tao, and H. Liu, "Bandwidth gain from mobile edge computing and caching in wireless multicast systems," *IEEE Trans. Wireless Commun.*, vol. 19, no. 6, pp. 3992–4007, Jun. 2020.

[12] M. J. Salehi, S. A. Motahari, and B. H. Khalaj, "On the optimality of 0–1 data placement in cache networks," *IEEE Trans. Commun.*, vol. 66, no. 3, pp. 1053–1063, Mar. 2018.

[13] S. P. Shariatpanahi, S. A. Motahari, and B. H. Khalaj, "Multi-server coded caching," *IEEE Trans. Inf. Theory*, vol. 62, no. 12, pp. 7253–7271, Dec. 2016.

[14] S. P. Shariatpanahi, G. Caire, and B. H. Khalaj, "Physical-layer schemes for wireless coded caching," *IEEE Trans. Inf. Theory*, vol. 65, no. 5, pp. 2792–2807, May 2019.

[15] A. Tölli, S. P. Shariatpanahi, J. Kaleva, and B. H. Khalaj, "Multi-antenna interference management for coded caching," *IEEE Trans. Wireless Commun.*, vol. 19, no. 3, pp. 2091–2106, Mar. 2020.

[16] A. Tolli, S. P. Shariatpanahi, J. Kaleva, and B. Khalaj, "Multicast beamformer design for coded caching," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Jun. 2018, pp. 1914–1918. [Online]. Available: https://ieeexplore.ieee.org/document/8437354/

[17] E. Lampiris, A. Bazco-Nogueras, and P. Elia, "Resolving the feedback bottleneck of multi-antenna coded caching," *IEEE Trans. Inf. Theory*, vol. 68, no. 4, pp. 2331–2348, Apr. 2022.

[18] S. Mohajer and I. Bergel, "MISO cache-aided communication with reduced Subpacketization," in *Proc. IEEE Int. Conf. Commun.*, 2020, pp. 1–6.

[19] Y. Cao, M. Tao, F. Xu, and K. Liu, "Fundamental storage-latency tradeoff in cache-aided MIMO interference networks," *IEEE Trans. Wireless Commun.*, vol. 16, no. 8, pp. 5061–5076, Aug. 2017.

[20] Y. Cao and M. Tao, "Treating content delivery in multi-antenna coded caching as general message sets transmission: A DoF region perspective," *IEEE Trans. Wireless Commun.*, vol. 18, no. 6, pp. 3129–3141, Jun. 2019.

[21] M. J. Salehi, H. B. Mahmoodi, and A. Tölli, "A low-subpacketization high-performance MIMO coded caching scheme," in *Proc. 25th Int. ITG Workshop Smart Antennas*, 2021, pp. 1–6.

[22] M. Salehi, M. NaseriTehrani, and A. Tölli, "Multicast beamformer design for MIMO coded caching systems," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, 2023, pp. 1–5.

[23] M. NaseriTehrani, M. Salehi, and A. Tölli, "Multicast transmission design with enhanced DoF for MIMO coded caching systems," 2023, *arXiv:2304.13827*.

[24] X. Yang et al., "Communication-constrained mobile edge computing systems for wireless virtual reality: Scheduling and tradeoff," *IEEE Access*, vol. 6, pp. 16665–16677, 2018.

[25] T. Dang and M. Peng, "Joint radio communication, caching, and computing design for mobile virtual reality delivery in fog radio access networks," *IEEE J. Sel. Areas Commun.*, vol. 37, no. 7, pp. 1594–1607, Jul. 2019.

[26] H. B. Mahmoodi, M. J. Salehi, and A. Tolli, "Non-symmetric coded caching for location-dependent content delivery," in *Proc. IEEE Int. Symp. Inf. Theory*, 2021, pp. 712–717.

[27] H. B. Mahmoodi, M. Salehi, and A. Tölli, "Asymmetric coded caching for multi-antenna location-dependent content delivery," 2022, *arXiv:2201.11611*.

[28] H. B. Mahmoodi, M. J. Salehi, and A. Tolli, "Non-symmetric multi-antenna coded caching for location-dependent content delivery," in *Proc. IEEE Int. Conf. Commun.*, 2022, pp. 5165–5170.

[29] H. B. Mahmoodi, M. Salehi, and A. Tölli, "Multi-antenna coded caching for location-aware content delivery," 2023, *arXiv:2305.06858*.

[30] "Extended reality (XR) in 5G, version 16.1.0," 3GPP, Sophia Antipolis, France, 3GPP Rep. 26.928, 2020.

[31] D. Kumar, S. K. Joshi, and A. Tölli, "Latency-aware highly-reliable mmWave systems via multi-point connectivity," *IEEE Access*, vol. 10, pp. 32822–32835, 2022.

[32] *LTE; Evolved Universal Terrestrial Radio Access (E-UTRA); User Equipment (UE) Radio Transmission and Reception*, 3GPP Standard TS 36.101, 2021.

[33] "Study on 5G new radio(NR) access technology, version 15.0.0," 3GPP, Sophia Antipolis, France, Rep. 38.912, May 2017.

[34] E. Dahlman, S. Parkvall, and J. Skold, *5G NR: The Next Generation Wireless Access Technology*. Amsterdam, The Netherlands: Academic, 2020.

[35] E. Lampiris and P. Elia, "Adding transmitters dramatically boosts coded-caching gains for finite file sizes," *IEEE J. Sel. Areas Commun.*, vol. 36, no. 6, pp. 1176–1188, Jun. 2018.

[36] Q. Yan, X. Tang, Q. Chen, and M. Cheng, "Placement delivery array design through strong edge coloring of bipartite graphs," *IEEE Commun. Lett.*, vol. 22, no. 2, pp. 236–239, Feb. 2018.

[37] M. Salehi, A. Tolli, S. P. Shariatpanahi, and J. Kaleva, "Subpacketization-rate trade-off in multi-antenna coded caching," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, 2019, pp. 1–6.

[38] M. Salehi and A. Tölli, "Multi-antenna coded caching at finite-SNR: Breaking down the gain structure," 2022, *arXiv:2210.10433*.

[39] M. J. Salehi, E. Parrinello, S. P. Shariatpanahi, P. Elia, and A. Tolli, "Low-complexity high-performance cyclic caching for large MISO systems," *IEEE Trans. Wireless Commun.*, vol. 21, no. 5, pp. 3263–3278, May 2022.

[40] M. J. Salehi, E. Parrinello, H. B. Mahmoodi, and A. Tolli, "Low-subpacketization multi-antenna coded caching for dynamic networks," in *Proc. Joint Eur. Conf. Netw. Commun. 6G Summit EuCNC/6G Summit*, 2022, pp. 112–117.

[41] M. Abolpour, M. J. Salehi, and A. Tolli, "Coded caching and spatial multiplexing gain trade-off in dynamic MISO networks," in *Proc. IEEE Workshop Signal Process. Adv. Wireless Commun. (SPAWC)*, 2022, pp. 1–5.

[42] M. Abolpour, M. Salehi, and A. Tölli, "Cache-aided communications in MISO networks with dynamic user behavior: A universal solution," 2023, *arXiv:2304.11623*.

[43] E. Parrinello, A. Ünsal, and P. Elia, "Fundamental limits of coded caching with multiple antennas, shared caches and uncoded prefetching," *IEEE Trans. Inf. Theory*, vol. 66, no. 4, pp. 2252–2268, Apr. 2020.

[44] E. Parrinello, P. Elia, and E. Lampiris, "Extending the optimality range of multi-antenna coded caching with shared caches," in *Proc. IEEE Int. Symp. Inf. Theory*, 2020, pp. 1675–1680.

[45] "Study on XR (extended reality) evaluations for NR, version 17.0.0," 3GPP, Sophia Antipolis, France, Rep. 38.838, 2022.

[46] "Study on XR enhancements for NR, version 1.0.0," 3GPP, Sophia Antipolis, France, Rep. 38.835, 2022.

[47] "New WID on XR enhancements for NR, work item description," 3GPP, Sophia Antipolis, France, Rep. RP-223502, 2022.

[48] 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.

[49] V. K. Goyal, "Multiple description coding: Compression meets the network," *IEEE Signal Process. Mag.*, vol. 18, no. 5, pp. 74–93, Sep. 2001.

[50] M. J. Salehi, A. Tolli, and S. P. Shariatpanahi, "Coded caching with uneven channels: A quality of experience approach," in *Proc. IEEE Workshop Signal Process. Adv. Wireless Commun. (SPAWC)*, 2020, pp. 1–5.

[51] Q. Lan et al., "What is semantic communication? a view on conveying meaning in the era of machine intelligence," *J. Commun. Inf. Netw.*, vol. 6, no. 4, pp. 336–371, Dec. 2021.

[52] C. Chaccour, W. Saad, M. Debbah, Z. Han, and H. V. Poor, "Less data, more knowledge: Building next generation semantic communication networks," 2022, *arXiv:2211.14343*.

[53] T. X. Vu, S. Chatzinotas, and B. Ottersten, "Edge-caching wireless networks: Performance analysis and optimization," *IEEE Trans. Wireless Commun.*, vol. 17, no. 4, pp. 2827–2839, Apr. 2018.

[54] E. Lampiris, H. Joudeh, G. Caire, and P. Elia, "Coded caching under asynchronous demands," in *Proc. IEEE Int. Symp. Inf. Theory*, 2021, pp. 718–723.

[55] A. Tang, S. Roy, and X. Wang, "Coded caching for wireless backhaul networks with unequal link rates," *IEEE Trans. Commun.*, vol. 66, no. 1, pp. 1–13, Jan. 2018.

[56] K. Wan, M. Cheng, M. Kobayashi, and G. Caire, "On the optimal memory-load tradeoff of coded caching for location-based content," *IEEE Trans. Commun.*, vol. 70, no. 5, pp. 3047–3062, May 2022.

[57] M. A. Maddah-Ali and U. Niesen, "Decentralized coded caching attains order-optimal memory-rate tradeoff," *IEEE/ACM Trans. Netw.*, vol. 23, no. 4, pp. 1029–1040, Aug. 2015.

[58] S. Jin, Y. Cui, H. Liu, and G. Caire, "A new order-optimal decentralized coded caching scheme with good performance in the finite file size regime," *IEEE Trans. Commun.*, vol. 67, no. 8, pp. 5297–5310, Aug. 2019.

[59] E. Parrinello, A. Bazco-Nogueras, and P. Elia, "Fundamental limits of topology-aware shared-cache networks," 2023, *arXiv:2302.10036*.

[60] M. Dutta and A. Thomas, "Decentralized coded caching for shared caches," *IEEE Commun. Lett.*, vol. 25, no. 5, pp. 1458–1462, May 2021.

[61] M. Ji, G. Caire, and A. F. Molisch, "Fundamental limits of caching in wireless D2D networks," *IEEE Trans. Inf. Theory*, vol. 62, no. 2, pp. 849–869, Feb. 2016.

[62] H. B. Mahmoodi, J. Kaleva, S. P. Shariatpanahi, and A. Tolli, "D2D assisted beamforming for coded caching," in *Proc. IEEE Wireless Commun. Netw. Conf. (WCNC)*, 2020, pp. 1–6.

[63] H. B. Mahmoodi, J. Kaleva, S. P. Shariatpanahi, and A. Tölli, "D2D assisted multi-antenna coded caching," *IEEE Access*, vol. 11, pp. 16271–16287, 2023.

MOHAMMADJAVAD SALEHI (Member, IEEE) received the B.Sc., M.Sc., and Ph.D. degrees in electrical engineering from the Sharif University of Technology, Tehran, Iran, in 2010, 2012, and 2018, respectively. Since 2019, he has been a Postdoctoral Researcher with the Center for Wireless Communications, University of Oulu, Finland. His research interests include coded caching and multiantenna communications.

KARI HOOLI received the D.Sc. (Tech.) degree in electrical engineering from the University of Oulu, Finland, in 2003. He is currently a Distinguished Member of the Technical Staff, Nokia Standards, Oulu. Before joining Nokia, he worked with the Centre for Wireless Communications, University of Oulu and visited the Centre for Communication Systems Research, University of Surrey, U.K. He holds numerous patents on 4G and 5G technologies and has authored several conference papers, journals, and book chapters. His research interests include physical-layer design and signal processing for wireless communications and cellular networks.

JARI HULKKONEN received the M.Sc.E.E. degree from the University of Oulu, Finland, in 1999. He has been working with Nokia since 1996. He started his career at Nokia in GSM/EDGE research and standardization projects. Since 2006, he has been leading radio research with Nokia Standards, Oulu, where he is currently the Radio Research Department Head with a focus on the 5G New Radio evolution. He has more than 30 patents/patent applications in 2G–5G technologies as well as several publications and book chapters.

ANTTI TÖLLI (Senior Member, IEEE) received the D.Sc. (Tech.) degree in electrical engineering from the University of Oulu in 2008. He is a Professor with the Centre for Wireless Communications, University of Oulu, Finland. From 1998 to 2003, he worked with Nokia Networks as a Research Engineer and the Project Manager in Finland and Spain. In May 2014, he was granted a five-year (2014–2019) Academy Research Fellow post by the Academy of Finland. He has authored numerous papers in peer-reviewed international journals and conferences and several patents. His research interests include radio resource management and transceiver design for broadband wireless communications, with a special emphasis on distributed interference management in heterogeneous wireless networks.