Challenges in Networking to Support Augmented Reality and Virtual Reality

Cedric Westphal
Huawei Technologies & University of California, Santa Cruz
Santa Clara, CA
Email: cedric.westphal@huawei.com

Abstract—We consider the network implications of virtual reality (VR) and augmented reality (AR). While there are intrinsic challenges for AR/VR applications to deliver on their promise, their impact on the underlying infrastructure will be undeniable. We look at augmented and virtual reality and consider a few use cases where they could be deployed. These use cases define a set of requirements for the underlying network. We take a brief look at potential network architectures. We then make the case for Information-centric networks as a potential architecture to assist the deployment of AR/VR and draw a list of challenges and future research directions for next generation networks to better support AR/VR.

I. INTRODUCTION

Augmented Reality and Virtual Reality are becoming common place. Facebook and YouTube have deployed support for some immersive videos, including 360 videos. Many companies, including the aforementioned Facebook, Google, but also Microsoft and others, are offering devices to view virtual reality, ranging from simple mechanical additions to a smart phone, such as Google Cardboard to full fledged dedicated devices, such as the Oculus Rift. While the first commercial deployments, such as Google Glass for augmented reality, were inauspicious, this new wave of products promises an enhanced experience to users.

Current networks however, are still struggling to deliver high quality video streams. 5G Networks will have to address the challenges introduced by the new applications delivering augmented reality and virtual reality services. The 5G white paper [2] mentions augmented reality, 3D-video and pervasive video as use cases for dense urban networks. However, it is unclear that without architectural support, it will be possible to deploy such applications.

Most surveys of augmented reality systems (say, [21]) focus on the actual displays and potential applications but mostly ignore the potential underlying network issues. We attempt to present some of these issues in this paper.

We also intend to explain how an Information-Centric Network architecture is beneficial for AR/VR. Information-Centric Networking has been considered for enhancing content delivery by adding features that are lacking in an IP network, such as caching, or the requesting and routing of content at the network layer by its name rather than a host’s address. It has been considered in the context of IoT [17], like AR/VR one of the potential drivers of traffic in the future Internet.

This paper is organized as follows. In Section II we attempt to frame the discussion by defining AR/VR and describing a non-exhaustive list of potential use cases. Then in Section III we list some requirements for the networking infrastructure. In Section IV we consider the network architecture and in particular the applicability of Information-Centric Networks. In Section V, we list some research challenges before offering some concluding thoughts in Section VI.

II. AUGMENTED REALITY AND VIRTUAL REALITY

We provide definitions of virtual and augmented reality (see for instance [21]):

Augmented Reality: an AR system inserts a virtual layer over the user’s perception of the real objects, which combines both real and virtual objects in such a way that they function in relation to each other, with synchronicity and the proper depth of perception in three dimensions.

Virtual Reality: a VR system places the user in a synthetic, virtual environment with a coherent set of rules and interactions with this environment and the other participants in this environment.

Virtual reality is immersive and potentially isolating from the real world, while augmented reality inserts extra information onto the real world.

For the purpose of this article, we restrict ourselves to the audio-visual perception of the environment (even though haptic systems may be used) as a first step. Many of the applications of augmented and virtual reality similarly start with eyesight and sounds only.

Most of the AR/VR we consider here focuses on head-mounted displays, such as Oculus Rift or Google Cardboard.

There are obvious observations derived from these descriptions of virtual and augmented reality. One is that virtual reality only really needs a consistent set of rules for the user to be immersed into it. It could theoretically work on a different time scale, say where the reaction to motion is slower than in the real world\(^1\). Further, VR only needs to be self-consistent, and does not require synchronization with the real world.

As such, there are several levels of complexity along a reality-virtuality continuum. For the purpose of the networking infrastructure, we will roughly label them as 360/immersive video, where user is streaming a video stream with a specific

\(^1\) At a cost of potential motion sickness from the user...
A. Use Cases

The 5G White Paper [2] describes use case for 5G which include augmented reality and what it denotes as pervasive videos. This is part of the Broadband Access in Dense Area use case. 3D services are expected to play a crucial part in 2020 and beyond. For this use case, the 5G white paper quotes 300Mbps of downlink, 50Mbps of uplink, an end-to-end latency of 10ms, and a mobility that is on demand between 0 and 100km/h, with thousands of users per km².

For AR/VR specifically, there is a range of scenarios with specific requirements. We denote a few below, but make no claim of exhaustivity: there are plenty of other applications.

1) Office productivity, personal movie theater: This is a very simple, canonical use case, where the headmounted device is only a display for the workstation of the user. This has little networking requirements, as all is collocated and could even be wired. For this reason, it is one of the low hanging fruits in this space. The main issue is of display quality, as the user spends long hour looking at a screen, with a resolution, a depth of perception, and a reactivity of the headmounted display that should be comfortable for the user.

2) Retail, Museum, Real Estate, Education: The application recreates the experience of being in a specific area, such as a home for sale, a classroom or a specific room in a museum. This is an application where the files may be stored locally, as the point is to replicate an existing point of reference, and this can be processed ahead of time.

Issues become then how to move the virtual environment onto the display. Can it be prefetched ahead of time; can it be distributed and cached locally near the device; can it be rendered in the device?

3) Sports: This attempts to put the user in the middle of a different real environment, as in the previous case, but adds to it several dimensions: that of real time, as the experience must be synchronized with a live event; that of scale, as many users may be attempting to participate in the experience simultaneously.

These new dimensions add some corresponding requirements, namely how to distribute live content in a timely manner that still corresponds to the potentially unique viewpoint of each of the users; how to scale this distribution to a large number of concurrent experiences. The viewpoint in this context also may impose different requirements, if it is that of a player in a basketball game, or that of a spectator in the arena. For instance, in the former case, the position of the viewpoint is well defined by that of the player, while in the latter, it may wildly vary.

4) Gaming: Many games place the user into a virtual environment, from Minecraft to multi-user shooter game. Platform such as Unity 3D allow creation of virtual worlds. Unlike the previous use case, there are now interactions in between the different participants in the virtual environment. This require communication of these interactions in between peers, and not just from a server onto the device. There are issues of consistency across users and synchronization issues.

5) Maintenance, Medical, Therapeutic: There exist a few commercial products where the AR is used to overlay instructions on top of some equipment so as to assist the agent in performing maintenance. Surgical assistance may fall in this category as well.

The advantage of a specific task is that it facilitates the pattern recognition and the back-end processing as it is narrowed down. However, the requirements to overlay the augmented layer on top of the existing reality puts stringent synchronization and round-trip time requirements, both on the display and on the sensors capturing the motion and position.

6) Augmented maps and directions, facial recognition, teleportation: The more general scenario of augmented reality does not focus on a specific, well defined application, but absorbs the environment as observed by the user (or the user’s car or the pilot’s plane, if the display is overlaid on a windshield) and annotates this environment, for instance to specify directions. This includes recognizing patterns and potentially people with the help of little context beyond the position of the user. Another main target of AR is telepresence, where a person in a remote location could be made present, as if in another location, say with others in the same conference room. Teleportation plus the display of the workstation of a user (as in the first scenario above) may allow remote collaboration on entreprise tasks.

III. NETWORK IMPACT & REQUIREMENTS

Based upon the previous use cases, we can specify some requirements on the network architecture.

The typical network architecture would be a hierchical server architecture, where functions are distributed to the edge of the network to ensure responsiveness and to accomodate somewhat limited populations of users.

The client must be equipped with a channel towards the server hierarchy to signal its position and what view it is observing in the virtual environment (VE). It also must be equipped with a channel in the other direction to receive the data pertaining to this VE. The client must be able to communicate directly with other clients in peer-to-peer application and shared VEIs.
A. Network Requirements

The survey [18] focuses on mobile and wireless technologies for AR/VR. However, it only considers the networking technologies up to 3G, with relatively low bandwidth and relatively high round-trip times (RTT) compared to the 5G targets. The improvement in wireless technologies since this report was published in 2008 have allowed new opportunities for AR/VR.

Macedonia et al. [16] draw a list of communication issues; distribution (unicast, multicast or broadcast); latency (lag, jitter); reliability (Acks, NAcks) and bandwidth. These all imply some corresponding requirements.

Bandwidth: The bandwidth requirements are unlikely to be supported in 5G for some applications of AR/VR for the resolution and responsiveness that is necessary for extended use. For instance, Bo Begole [4] calculates that human can process 5.2Gbps of data based on the physical characteristics of human perception. This is calculated based upon the ability to distinguish 200 dots per degree within the typical human foveal field of view, with at least 150 degrees horizontally and 120 degrees vertically. Add to this 30 frames per second and some compression ratio, and you get to 5.2Gbps. Recall that the download figure in the “dense urban broadband” use case in 5G targets 300Mbps: the bandwidth needed for full VR is twenty times more than the target bandwidth.

Delay: [1] cites a requirement of millisecond end-to-end RTT with high reliability for augmented reality. This is a factor 10x compared to the 5G target of 10ms. There are two types of delay in AR/VR. One is the delay of the application being transmitted. For instance, for a telepresence application, the other person in the communication needs to respond to us within 100ms. This is the same delay as in a videoconference today. The other is the delay perceived as the user changes point of view. For instance, if the user looks to one side, then to the other, the device has to be responsive in order to avoid the user’s sickness/nausea. This delay has to be less than 10ms. It includes all the steps, including the rendering on the screen, which for current LCD technology takes from 10ms to 20ms. Pico LDP display technology is much faster.

As a consequence, the processing of the images is physically impossible from a long distance. One cannot have the server on one coast of the US and the user on the other, and expect the application to be hosted remotely. The processing should be either on the user’s device, or on a server at the network edge. This imposes a distributed architecture.

Computation requirements: it is difficult to estimate the computation requirements for AR/VR; however, one interesting data point can be found in [14]. [14] considers an Intelligent Personal Assistant (IPA) system, like Siri on an Apple iphone, or the “Ok, Google” feature on Android. The steps in such a system are: doing speech-recognition on the request, recognizing the query (classified as question/action), processing the query by looking up some database, and returning an answer. This follows the steps of an AR system, where the vocal command is replaced by a video view from the user’s device, which would then be processed; the features extracted and the patterns identified; the question (say, what are the streets and buildings in the view) answered by looking up a database; the answer returned and then layered onto the view.

From these steps, the complexity of the IPA system is much less than that of AR. The features/semantic extraction comes from snippets of voice, not a video stream; the result can be processed in the order of seconds, and the answer does not have to be rendered and overlayed in real time; further, IPA tackles one query at a time, while AR would continuously generate new perspectives and therefore new tasks to resolve. [14] estimates what it calls the scalability gap as a factor 100x (without optimization) if IPA queries replaced web searches. This means the worldwide number of DCs would need to be scaled up by 100 times to satisfy IPA queries. We conjecture that this number of 100x is conservative even for an optimized AR system.

One architectural design goal is therefore to place the processing for such a system as close to the edge of the network as possible, as each backend system will not be able to accommodate the same volume of requests as for today’s DCs. Another consequence is that only an AR system that is well integrated into the network infrastructure can deliver the desired performance.

IV. NETWORK ARCHITECTURE

From the above discussion, it is clear that current AR/VR applications are requiring network resource that are difficult to meet. Therefore, the network infrastructure needs to be re-evaluated in order to support such apps.

[10] looked at the impact of network topology on the performance of scalable multi-user virtual environments. Peer-to-peer flat topologies, and client-server hierarchical topologies are considered.

NetEffect [8] is a specific embodiment of a virtual environment which is partitioned into different communities, each one of which is allocated to a specific server. The client visiting the corresponding part of the VE is migrated to the proper server.

Mobile Edge Computing (MEC) [15] presents a network architecture for 5G which supports virtual reality. There has been some work on video distribution using edge computing [12] [25]. The MEC server holds AR data and AR object caches for high bandwidth/low latency content delivery.

A. Functions and APIs

While it is possible that AR/VR will happen purely as an overlay play, we contend that properly placed functions and properly defined APIs will alleviate the potential networking issues.

Virtual reality needs to transmit a representation of a VE. One potential way is to transmit the whole representation, and let the device select which specific view to display. However, this is an inefficient way to use the network, as most of the
representation is never observed and it is not necessary to transmit it.

In the specific sub-case of 360 video, [19] looked at ways to minimize the amount of transmitted data. It does it through two mechanisms: one is a representation of the VE that can be packetized into views, such that only the required views need to be transmitted.

The other is a prediction mechanism of what direction the user will be looking at, what her position will be, and using this prediction, only the predicted views are transmitted (plus some buffer around it to correct for prediction errors), or only the predicted landscape is pre-processed.

Integrating AR/VR with the network infrastructure would require to open up APIs towards these two functions (namely an encoding mechanism and a prediction mechanism) such as to minimize the bandwidth.

Integrating pre-processing of augmented reality in the infrastructure (for instance, at the access point serving the user) would go a great way towards solving the RTT issues. It is unclear that an OTT player could reach this level of server placement.

B. Information-Centric Network Architecture

We now turn our attention to the potential benefits that Information-Centric Networks can bring to the realization of AR/VR.

The abstractions offered by an ICN architecture are promising for video delivery [5] [13] [25]. RFC7933 [22] for instance highlights the challenges and potential of ICN for adaptive rate streaming. As VR in particular may encompass a video component, it is natural to consider ICN for AR/VR.

There is a lot of existing work on ICN (say, caching [23] [24] or traffic engineering [3] [20] [7]) which could be applied to satisfy the QoS requirements of the AR/VR applications, when possible.

One of the key benefits from ICN is the native support for multicast. For instance, [16] quotes: "if the systems are to be geographically dispersed, then highspeed, multicast communication is required." Similarly, [9] states that: "Scalability is achieved by making extensive use of multicast techniques and by partitioning the virtual universe into smaller regions."

In the sport use case, many users will be participating in the same scene. They will have potentially distinct point of views, as each may look into one specific direction. However, each of these views may share some overlap with the others, as there is a natural focus point within the event (say, the ball in a basketball game).

This means that many of the users will request some common data and native multicast significantly reduces the bandwidth and in the case of ICN, without extra signaling.

Further, the multicast tree should be adhoc, and dynamic to efficiently support AR/VR. Back in 1995, [11] attempted to identify the visual interactions in between entities representing users in a VE so as to "reduce the number of messages required to maintain consistent state among many workstations distributed across a wide-area network. When an entity changes state, update messages are sent only to workstations with entities that can potentially perceive the change i.e., ones to which the update is visible." [11] was able to reduce the number of messages processed by client workstations by a factor of 40.

It is unclear that ICN can assist in identifying which workstations (or nowadays, which users) may perceive the status update of another user (but naming the data at the network layer may help). Nonetheless, the multicast tree to reach the set of clients that would require an update is dynamically modified and the support for multicast in ICN definitely supports this dynamic behavior.

The caching feature of ICN allows prefetching of data near the edge some of the more static use cases; further, in the case of multiple users sharing a VE, the caching allows to perform the content placement phase for some users at the same time as the content distribution phase of others, thereby reducing bandwidth consumption.

V. Research Challenges

As the literature quoted in this paper indicate, a lot of the research work was performed while the underlying network were unable to support AR/VR properly. As we have observed, the processing, bandwidth and delay requirements will not be met in the WAN with 5G without some compromise on the part of the end users. However, as early cellular networks have proved for voice calls, users are able to be satisfied with less than stellar QoS as long as a new value proposition is delivered.

Some research challenges are therefore:

- to identify the network architecture that can deliver AR/VR capability in 5G networks; in particular, which functions are necessary, and how to integrate this functions within the network architecture;
- to identify the interfaces required from AR/VR to the network; in particular, can AR/VR function purely as overly, or should AR/VR require some help from the infrastructure for caching, multicasting, traffic engineering, QoS, etc.
- to identify the proper naming semantics to expose at the network layer enough information to allow the sharing of data in between several VE sessions;
- to identify the coding of the VE so that a VE can be packetized into multiple view "cells" units that can be recomposed into a VE sessions, and that can be shared in between different sessions;
- to characterize the motion prediction and the corresponding network protocol; to assess whether this information needs to be shared with the infrastructure;
- to identify rate adaption mechanism for AR/VR, similar to the rate adaptive mechanisms in video distribution;
- to identify caching policies for AR/VR content;
- to specify QoS on the fly, using SDN or similar control tools; to specify what an SDN controller needs to know about an AR/VR application;
• to characterize the reliability requirements of AR/VR sessions at the network layer;
• to consider possible transport protocols that are designed for AR/VR; transport protocols could be adapted to the different channels from the user (say, for updating its position with very low latency; or for receiving a remote image, potentially with less stringent delay);
• to identify security mechanisms;

There are of course many other challenges, we hope this paper triggers thoughts on which other ones are worthy of further study. This area is ripe for investigation.

VI. CONCLUSION

We considered the impact of AR/VR on the networking infrastructure. With 5G, networks and in particular wireless networks, will be one step closer from ubiquitously supporting AR/VR. We considered a few use cases, and looked at how these scenarios created new requirements on the underlying networks. The support for multi-cast has been requested for AR/VR applications for a long time: Information-Centric Networks are believed to provide significant benefits to better support AR/VR.

REFERENCES