Support Precise Latency for Network Based AR/VR Applications with New IP

Lijun Dong, Lin Han, Richard Li
{liju.dong, lin.han, richard.li}@futurewei.com

Futurewei Technologies Inc., 2330 Central Expressway, Santa Clara, CA, USA, 95050

Abstract. The emerging Augmented Reality (AR), Virtual Reality (VR) and Holographic applications have brought up a lot of challenges to technologies such as information display, image processing, fast computing and networking. This paper gives a quantified analysis on the latency requirement that AR and VR impose to networking. Most importantly, the paper discusses on how to use New IP, an advanced data packet framework to support the precise low latency requirement.

Keywords: AR, VR, networking, latency budget, precise latency constraint, New IP, contract spec, metadata, holographic application, hologram.

1 Introduction

The multimedia has been evolved from the 2-D audio/video in 4K/8K resolution to AR/VR video. Augmented Reality (AR) is a live direct or indirect view of a physical, real-world environment whose elements are augmented by computer-generated sensory input such as sound, video, graphics or GPS data. Virtual Reality (VR) is a computer technology that uses software-generated realistic images, sounds and other sensations to replicate a real environment or an imaginary setting, and simulates a user's physical presence in this environment to enable the user to interact with this space. A VR viewport delimits the scene horizontally from the viewport center, whose angle can normally range 120 degrees. In order to ensure good immersion, a displayed viewport's pixels need to at least have 4K resolution. Therefore, the resolution of the full 360-degree scene is at least 12K.

AR/VR technologies have enormous potential in many different fields, such as entertainment, remote diagnosis, and remote maintenance, etc. Holograms, haptics, and other sensory data will further provide immersive and feel-like-real user experience, with which the real and virtual world to the users will be extensively blurred. In a hologram, the same object or scene is captured from different angles. A different image depicting the same object or scene will be seen by a viewer from different viewpoints, depending on the relative postion of the reviewer's eyes.

Powerful cloud capabilities have improved the VR user experience and reduced device cost, promoting the evolution of VR from local to cloud-based VR. Cloud VR can make full use of the distributed computing capabilities of many-core servers, Graphics Processing Unit (GPU) clusters, as well as latest rendering and artificial intelligence (AI) technologies. **Figure 1.** shows an example of cloud VR - virtual concert. In the virtual concert, the musicians could be performing in different places in the world, while the audience could be sitting on the beach enjoying the concert as if he is present in front of the stage with those musicians' 3D projections. The holographic data is represented through use of point clouds consisting of volumetric data in

EAI MOBIMEDIA 2020, August 27-28, Harbin, People's Republic of China

DOI 10.4108/eai.27-8-2020.2294291

a conceptual three-dimensional box. The large volumetric data needs to be streamed through the Internet to the end user such that the redering of the interested object/scene's image from any 360-degree viewing angle can be achived. We call the AR/VR applications that involve large volumetric data streaming as network-based AR/VR in the following sections of the paper.

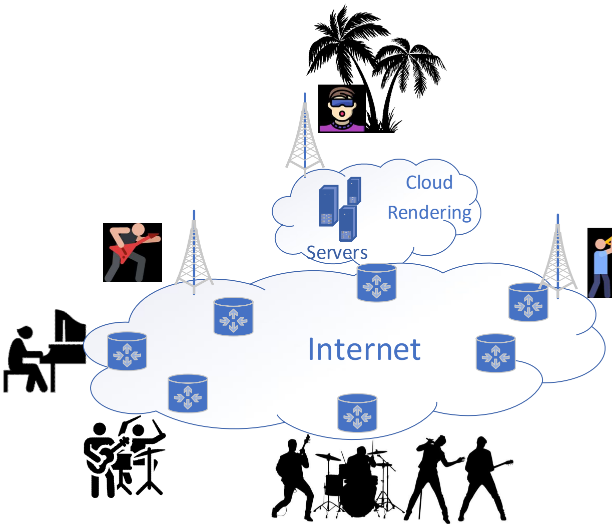

Fig. 1. Virtual concert.

"VR sickness" [1] already shows in AR/VR applications if the information about the virtual environment received by the human brain is not always consistent due to data transmission lagging. In the network-based AR/VR, sufficient realism inevitably requires both extremely low latency and high bit rate.

The paper focuses on providing a quantified analysis on transport latency budget in supporting network-based AR/VR applications. Most importantly, in the paper the authors discuss on how to achieve precise latency budget by leveraging a novel and thriving data packet framework called New IP.

2 Latency Budget

Latency is the most important performance metric for AR/VR, holographic streaming. Motion to Photon (MTP) is defined as the time needed for user movement to be fully reflected on a display screen. A high MTP latency causes sickness and nausea. When a user wearing VR headset makes a movement, the mind expects the display on the screen is also updated promptly and appropriately to reflect the movement. When MTP latency is high, the display also fails to show the user's movement, then the user can be disoriented and feel sick, resulting in very poor and intolerable VR experience. It often considers that the MTP latency less than 20 ms is necessary to convince your brain that you are presented in an augmented or simulated world. Some research even shows that the MTP latency must be smaller than 17ms [2] for sensitive users. Latency greater than 20 ms not only degrades the visual experience, but also tends to result in VR sickness, which is also known as cyber-sickness. It is caused by a sensory mismatch or conflict to the signals the balance system is sending to the brain. Taking the analogy to riding in a car, the vestibular system tells the brain that you are moving, but the proprioceptive system indicates that you are sitting still. You may experience car sickness if your visual system may be getting different signals depending on what you are looking at, and confuses your brain about your movement since you cannot predict exactly when you are going to slow down, speed up or turn as a passenger. VR sickness can be minimized by keeping MTP latency below the threshold, above which humans can detect the lag between the visual input and self-movement.

Table 1. Current and projected latency in network based AR/VR

Latency	Current value (ms)	Projected value (ms)
T1	1	1
T2	11	2
T3	110 to 1000	5
T4	0.2 to 100	?
T5	5	5
T6	1	0.01
MTP	130 to 1180	13+?

The network-based AR/VR involves many more factors from one end to another end. The major components of E2E delay in network-based AR/VR include: T1: Sensor detection and action capture; T2: Computation for region of interest (ROI) processing, rendering and encoding; T3: Group of Pictures (GOP) [3] framing and streaming; T4: Network transport; T5: Terminal decoding; T6: Screen refresh.

Table 1. shows the current and projected values for each component of end-to-end latency. If we anticipate that the technology development and advancement would bring down the latency of some components, such as reducing the latency caused by ROI processing, rendering and encoding (T2) to 2 ms, GOP framing and streaming (T3) to 5ms by using improved parallel hardware processing, and screen refresh latency (T6) to 0.01ms by using OLED, etc., then the budget for the round trip network transport delay (T4) will be around 5 to 7ms.

We can see that MTP latency is currently much greater than 7ms. The network transport latency is comprised of physical propagation delay and switching/forwarding delay at each network device.

- The physical propagation delay: This is the delay caused by the speed limit of signal transmitting in physical media. Taking the fiber as an example, the optical transmission rate cannot exceed the light speed, i.e. 300km/ms in free space. However, light travels slower the fiber optic core because the refractive indexes of light are different in free space and in the glass. In normal optical fiber, the light speed is about 200km/ms [4]. In order to reduce the physical propagation delay, the physical distance between user and AR/VR server needs to be limited. The deployment of AR/VR server should be close to user as much as possible.

- The switching/forwarding delay: This delay normally is much more than the physical propagation delay, which can vary from 200us to 200ms at each hop.

We cannot expand the physical scope of an AR/VR application beyond the speed-of-light limit. However, we can ensure that application processing and transport related latencies do not significantly reduce this limited scope. As a rule of thumb, they should consume no more than 5-10% (1-2ms) of this 20ms budget, and preferably less.

3 Support Precise Latency Budget by New IP

3.1 What is New IP?

New Internet Protocol (New IP) [5][6] [8] has been proposed to overcome the three major issues that are caused by the fixed structure of the IP packets in the current Internet, e.g. statistical multiplexing, best-effort paradigm, and an IP address-based reachability. New IP is a data plane technology that defines a new network datagram format, its specification, and corresponding capabilities in the network nodes. The New IP datagram format is shown in **Figure 2.**, which includes 3 components, namely, a) addressing evolution, b) the contract inclusion, and c) the payload extension.

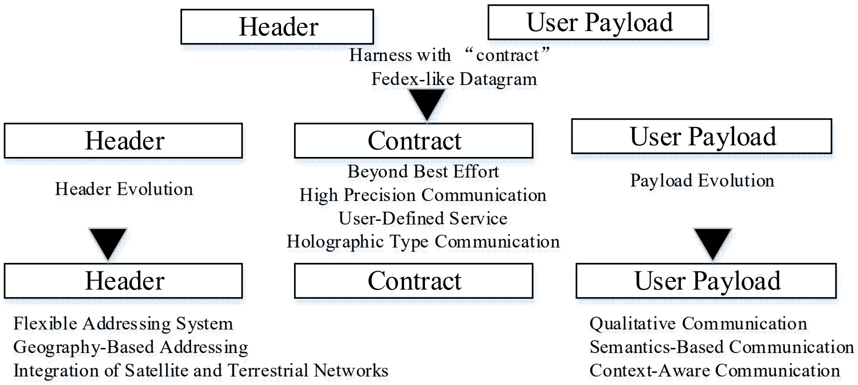
Header
User Payload
Harness with "contract"
Fedex-like Datagram
Header
Contract
User Payload
Header Evolution
Beyond Best Effort
High Precision Communication
User-Defined Service
Holographic Type Communication
Payload Evolution
Header
Contract
User Payload
Flexible Addressing System
Geography-Based Addressing
Integration of Satellite and Terrestrial Networks
Qualitative Communication
Semantics-Based Communication
Context-Aware Communication

Fig. 2. New IP.

The *New IP Address* (Shipping Spec) evolution aims to replace the current fixed type of addressing in order to provide flexibility to include all types of addresses and fit different reachability scenarios. The New IP shipping specification is backward compatible with the existing address schemes (e.g., IPv4 and IPv6).

The *New IP Contract* (Contract Spec) inclusion provides a series of apparatus to enable a large selection of network capabilities, their functioning and regulative control at the finest packet-level granularity. Contracts carry specific requirements and parameters associated with time-engineered services for media services as we discussed in the previous section of this paper. **Figure 3.** shows the contract structure. Basically contract is composed of multiple contract clauses, each of which is a combination of Event, Condition, Action and Metadata. Action describes how New IP nodes should treat the packet when certain designated event/condition is met. The metadata is a set of parameters that are associated with the actions or applications.

The *New IP Payload* (Payload Spec) associates semantics to the data payload. New IP payload provides options to the receiver to consume any residual information in the payload while allowing the network to drop portions of the payload when congestion occurs. This type of communication is named as Qualitative Communication [7], which helps to mitigate re-transmission overheads and delays when faced with slow or congested conditions.

```
<Contract> := <Contract clause>
        | <Contract clause> AND <Contract>

<Contract clause> := <Contract ECA>
        | <Contract ECA> OR <Contract clause>

<Contract ECA> := <Event, Condition, Action>
        | <Metadata>
        | <Event, Condition, Action> <Metadata>
        | <Action>
        | <Action><Metadata>
```

Fig. 2. New IP Contract.

3.2 How New IP Supports the Precise Latency Requirement?

The Contract Spec in New IP format is able to carry latency requirement to precisely support computational multiplexing approaches on the switches or routers, which provides much finer granularity of time assurance at packet level compared to the current statistical multiplexing.

For AR/VR or holographic data (later is called multimedia data in general), by inserting the precise latency requirement in the New IP contract clause (Metadata), the residual latency can be evaluated in a hop-by-hop basis by intermediate New IP node (switches or routers). We regard the packets with time constraints as latency-sensitive packets. A computational multiplexing scheduler at each New IP node is capable of determining the precise position in the outgoing queue of an output port for a latency-sensitive packet based on its latency constraint compared to other packets. With computational multiplexing, simultaneously arriving packets with latency budget that are intended to be forwarded on the same output port are scheduled based on their respective latency budgets. This capability is only possible when datagrams have flexible structure as designed in New IP to carry their latency budget because control plane methods are not designed to handle per packet requirements.

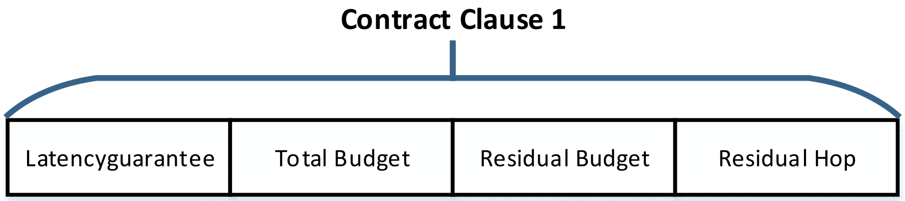

Fig. 3. Contract Clause 1.

In order to ensure the extreme low latency with precise budget, two contract clauses are designed for the multimedia data packets. The first contract clause is shown in **Figure 3.**, the action is $Latencyguarantee$, which instructs the intermediate New IP nodes to perform computational multiplexed scheduling of the packet. The metadata includes: (1) `total budget` is the end-to-end latency constraint between the time when the sender sends out the packet and the time when the packet reaches the receiver, which is set to be at most 2ms for cloud-based AR/VR applications; (2) `residual budget` is computed as the total budget subtracted by the elapsed time when the packet arrives at a New IP node; (3) `residual hop` indicates the number of hops between the current node to the receiver. One straightforward scheduling algorithm is to place the latency-sensitive packets in one prioritized queue per output port, and schedule them according to average per-hop residual budget, which is defined as residual budget divided by residual hop. This scheduling algorithm makes sure the packet with the smallest latency budget gets transmitted ahead of the packets with larger budgets. However, it has a fundamental fairness issue, that the packets with larger per-hop residual budget are being “starved” at the end of the queue. The researches on this topic are highly encouraged.

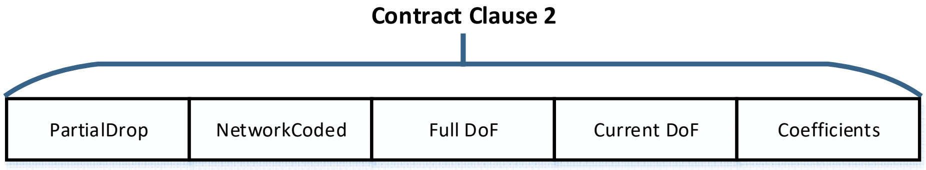

Fig. 4. Contract Clause 2.

The second contract clause is designed to instruct the intermediate New IP nodes to drop portions of packet payload when encountering network congestions instead of dropping the packet completely. The action is set to be $PartialDrop$. It is proposed in [10][11], the packet payload could be divided into multiple equally sized chunks, over which the random linear network coding is applied. When network congestion happens that requires dropping the latency-sensitive packet entirely (according to the current implementation in the Internet), the New IP node could remove/trim chunks from the tail of the packet payload as many as needed until the outgoing queue is able to retain the packet. In this way, the receiver is still able to obtain some parts of the packet payload, such that only the lost portions are retransmitted. In the contract clause 2, some metadata is also configured to assist $PartialDrop$ execution, which includes: (1) `NetworkCoded` is used to indicate that the packet payload chunks are applied with random linear network coding; (2) `Full DoF` indicates the complete number of degrees of freedom in order to decode the payload data chunks; (3) `Current DoF` indicates the degree of freedom of the remaining chunks in the payload. When a chunk is removed from the payload, the current DoF is subtracted by 1; (4) `Coefficients` contains the coefficients for the remaining coded chunks in the payload. When a chunk is removed from the payload, the corresponding coefficient is also deleted from the Coefficients metadata.

4 Conclusion

The network-based AR/VR and future holographic applications impose many new challenges to the networking technologies, especially the transport layer. The paper analyzes the maximum value that the network transport latency could have. The paper discusses one possible solution to address the precise latency requirement, which leverages the flexible, programmable New IP data packet framework. The paper also outlines the research directions and opportunities when leveraging New IP framework.

References

[1] Wikipedia, "Virtual Reality Sickness", 2016.

[2] K. Raaen, “Response Time in Games : Requirements and Improvements”, University of Oslo, PhD Thesis, 2016.

[3] Wikipedia, “Group of Pictures”, 2016.

[4] K. Miller, “Calculating Optical Fiber Latency”, 2012.

[5] FG-NET-2030, Sub group 2, “New Services and Capabilities for Network 2030: Description, Technical Gap and Performance Target Analysis.” ”https://www.itu.int/en/ITU-T/focusgroups/net2030/Documents/Deliverable/NET2030.pdf”, 2019.

[6] R. Li, K. Makhijani and L. Dong, “New IP: A Data Packet Framework to Evolve the Internet,” in submission to IEEE HPSR 2020.

[7] R. Li, K. Makhijani, H. Yousefi, C. Westphal, L. Dong, T. Wauters, and F. De Turck, “A Framework For Qualitative Communications Using Big Packet Protocol,” in NEAT’19: Proceedings Of The 2019 ACM Sigcomm Workshop On Networking For Emerging Applications And Technologie, pp. 22–28, ACM, 2019.

[8] R. Li, A. Clemm, U. Chunduri, L. Dong, and K. Makhijani, “A New Framework And Protocol For Future Networking Applications,” ACM Sigcomm Workshop on Networking for Emerging Applications and Technologies (NEAT 2018), pp. 637–648, May 2018.

[9] A. Clemm, T. Eckert, “High-Precision Latency Forwarding over Packet - Programmable Networks,” IEEE/IFIP Network Operations and Management Symposium, 2020.

[10] L. Dong, R. Li, “In-Packet Network Coding for Effective Packet Wash and Packet Enrichment,” 2019 IEEE Globecom Workshop on Future Internet Architecture, Technologies and Services for 2030 and Beyond.

[11] L. Dong, K. Makhijani, R. Li, “Qualitative Communication Via Network Coding and New IP,” in submission to IEEE HPSR 2020.