Deploying a Low Latency Service on the Cable Edge

Virtual Reality Streaming

A Technical Paper prepared for SCTE by

Omar Alami
Principal Engineer I
Charter Communications
6399 S Fiddlers Green Cir #100, Greenwood Village, CO 80111
720-699-3433
Omar.Alami@Charter.com

Table of Contents

Title **Page Number**

List of Figures

Title **Page Number**

List of Tables

Title **Page Number**

1. Introduction

In this paper, Charter's Emerging Technology group will share its findings from the field trial of a low-latency virtual reality (VR) gaming service built on its edge computing premises. High-powered virtualized servers were configured for streaming interactive VR content to Spectrum Internet customers. Key metrics (latency, jitter, packet loss) were used to evaluate and benchmark network performance for high bandwidth, low latency services deployed at the Charter edge.

Recent trends in connectivity have accelerated use cases in which customers utilize high bandwidth but also require low latency. Emerging immersive use cases, such as virtual reality and augmented reality (AR), are inherently latency-sensitive due to the nature of being tethered to a user's head. This demand for low latency presents a significant challenge for hardware, especially in form factors that are consumer friendly and affordable to a mass market. Today, devices priced in line with gaming consoles suffer from low fidelity and devices with high fidelity require tethering to a gaming computer that usually costs thousands of dollars. The cable industry is poised to accelerate these technologies by offloading compute to high-powered servers on the network edge. With the industry's commitment to 10G service, higher levels of bandwidth can be utilized to enable experiences through the network that traditionally require high local compute.

To better understand the latency demands of these emerging use cases, Charter aimed to launch a field trial that tested our network more than typical existing use cases. Charter's Cloud Virtual Reality field trial is a system in which high-powered virtualized servers stream high fidelity interactive VR content from a regional data center (RDC) to customers on the Charter network. The aim of this field trial was to understand how an immersive use case that demands high bandwidth and low latency would perform in our existing network. Network and performance metrics were collected to draw conclusions on what performance our edge delivers in terms of latency, packet loss and jitter. Additional metrics of the client's connection were collected such as WiFi signal (RSSI), WiFi frequency and WiFi band. Metrics were collected on the server to understand how the infrastructure required to serve this experience performed, including render latency, encode latency, GPU utilization and CPU utilization. All together, these metrics establish key parameters and expectations on enabling low latency services through the network.

Any game streaming experience must consider the network latency as a key factor of the overall latency. Total input latency (i.e., the latency between a user action and the action being reflected in the display) is made up of the client-side latencies, the network latency and server-side latencies. When compared to a traditional gaming system, streaming a game also adds encode and decode latencies. A major challenge service providers face is how to utilize high powered compute to help make up for the additional latencies introduced by streaming. Techniques that utilize additional bandwidth were explored to help hide and minimize the impact of latency. These tradeoffs can be adjusted to allow for different levels of latency and fidelity depending on the use case and distance to users.

Learnings from this field trial help Charter understand how low latency services can perform on the network today and how this can change with future network enhancements such as Low Latency DOCSIS (LLD), Active Queue Management (AQM) and Low Latency, Low Loss and Scalable Throughput (L4S).

2. Emerging Network Powered Use Cases

The increasing availability of high-speed internet, including gigabit speeds, has enabled high bandwidth use cases such as 4K video streaming. Many households now have the capacity to handle multiple concurrent 4K streams without negatively impacting other internet applications within the household. Beyond 4K video streaming, it is unclear if 8K streaming will become a popular use case in the near term.

One domain poised to increase the demand for 4K and beyond streaming is VR and AR. Modern VR displays are 4K, or near 4K, and are an obvious use case for streaming 4K video. Head-mounted displays also have the ability to view 360-degree videos, in which 8K streaming is needed. This emerging use case is poised to continue growing with the increase of VR device sales led by the Meta Quest (formerly known as Oculus Quest) headsets and the announcement of the first headset from Apple. Use cases showcased by Apple include streaming high-definition content from streamers like Disney+, including immersive assets rendering within a user's environment. The sheer size of immersive video and content creates an obvious need for high bandwidth streaming.

It is possible that consumers determine 4K is indistinguishable enough from 8K on 2D TV panels so that 8K remains a niche use case. More immersive media consumption would be reserved for head-mounted displays. What would be next for the household TVs? One use case that is looking to challenge traditional TV displays are light field displays. Light field displays aim to create a field of light that shows distinct colors depending on the direction it is being viewed from. This is similar to how we take in light in the real world, in which light rays reflect off of objects and the angle of the rays determines the color and depth we perceive. A traditional TV display emits a color for each pixel regardless of the angle of the viewer. A light field display enables a sense of 3D and depth since the image is dynamic based on view angle. Modern light field displays are a major upgrade from traditional 3D TVs that failed to see adoption in peoples' homes. Some forms of light field displays require eye tracking, a similar problem to requiring 3D glasses. Mass adoption would require the same viewing convenience as traditional TVs. Displays need to not require special glasses or eye-tracking from cameras, while also covering a large field of view (FOV). Once this becomes the case, the potential for replacing household TVs will be very real. Supporting streaming to light field displays requires significantly more bandwidth than traditional TVs. Light field displays require multiple camera angles to create its image, so instead of streaming a single video stream, it will be equivalent to streaming up to 100 video streams, depending on the display and format. This could easily require hundreds of megabits per second to stream to a single display.

The need for more bandwidth will continue to grow, however providing high bandwidth alone to customers is not enough for many of these use cases, such as VR and AR. The network must be able to provide services with low latency at the same time. Low latency has always been desired for traditional use cases like live broadcasting. Viewers grow frustrated if they hear their neighbor cheer for a score in a sports game before their stream showed it to them. In the context of immersive use cases like VR and AR, low latency moves beyond just convenience and becomes a strict requirement. Streaming content to head-mounted displays requires low latency to prevent VR sickness, which can be caused by a delay in displaying updated images after head motion. Given these emerging use cases, the network will need to not only provide high bandwidth, but ensure it can be delivered with low latency.

3. Network Demands of the Future

Given these emerging use cases, Charter looks to find ways the network can be leveraged to enable better services. If consumer-grade light field displays were available today, could our existing networks support streaming to them? If the cost of hosting compute in the cable edge suddenly dropped, what kind of network-powered services could be supported over existing networks? Challenging the Charter network requires looking at emerging use cases and understanding the network demands of the future.

The recent increase in the adoption of VR headsets can be attributed to the affordability of standalone headsets. Traditional VR headsets are hundreds of dollars alone and require being tethered to a computer. The computer runs all of the content and requires a high-end GPU and CPU, often costing thousands of dollars. This has reserved VR setups for a niche market willing to spend the high upfront cost. Meta Quest headsets removed the tether and are available for only a few hundred dollars. While this significantly increased adoption, it came at the cost of visual fidelity. Standalone headsets suffer from tradeoffs between size, power, thermal dissipation and battery life. Keeping a headset compact and minimizing heat to the user's face means the GPU power that can be used in a standalone headset is limited. Traditional VR content can run at 4K+ with very high visual fidelity meeting the current standards of AAA video game graphics. Standalone VR content was forced to adopt lower fidelity graphics, with a lot of developers opting for content with simpler cartoon-like graphics rather than photorealistic content.

These tradeoffs offer a unique opportunity for service providers to leverage their networks. One tradeoff that standalone headsets haven't made compared to tethered headsets is screen quality and resolution. Despite having less compute power, the displays of standalone headsets can be just as good as tethered headsets. This means that the issue can be summarized as a compute location problem. Having compute outside of the headset could enable standalone headsets to display the same high-fidelity graphics of a tethered headset.

This opportunity is where Charter looked to understand the compute required to provide a service like this, as well as the network requirements. Could the existing Charter network support streaming video beyond 4K and with the low latency required for VR?

4. Cloud Virtual Reality

Cloud virtual reality is a service in which virtual reality content is rendered on a server and streamed to a head-mounted device. The most applicable example of a similar service is NVIDIA GeForce Now, which is a cloud gaming service that offers the ability to run high-end PC (personal computer) games on a cloud computer and have it streamed to a device that couldn't normally run those games. As shown in Figure 1 - Cloud Virtual Reality, Cloud VR works by having a standalone VR headset act as a thin client that sends out its inputs, including headset position/orientation and controller position/orientation with button states, to a cloud or edge server. The server, which consists of hardware capable of rendering high fidelity VR graphics, receives the input data, executes the content, renders and encodes frames sent back to the client. The client then decodes the video frames and displays it to the user. This concept can be achieved within a user's home if they have a computer that is capable of rendering PC VR and performs very well due to the low latency of a local network.

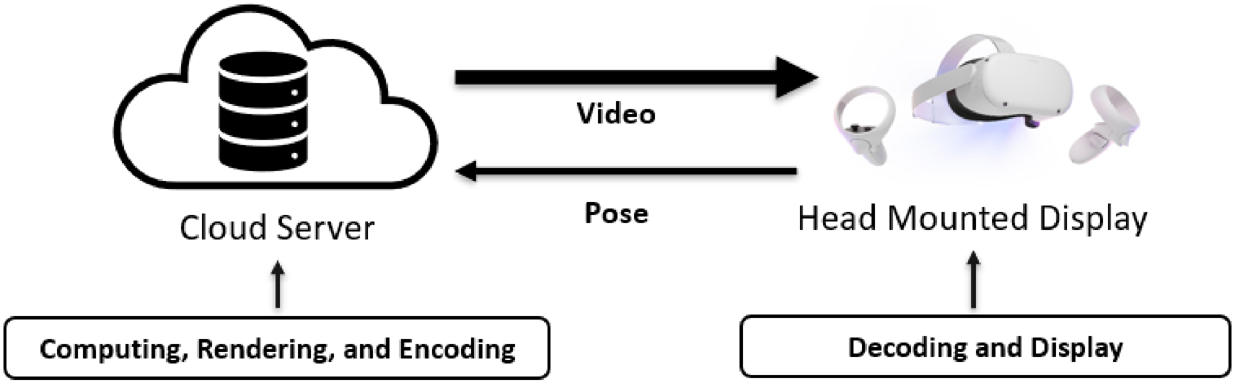

Figure 1 - Cloud Virtual Reality

Running a service like this from a traditional cloud is limited by latency. Cloud VR is a use case for utilizing an edge server, in which case the server is physically located closer to users than a traditional cloud server would be. The three main components to a service like this are the application, the infrastructure and the content. The application includes a client application and server application that communicate with each other. The infrastructure includes an edge server, consisting of servers with high-power GPUs and CPUs capable of running high fidelity VR games, and an edge network with efficient routing to users. The content includes the VR games and/or experiences that would run on the servers.

At first glance, introducing additional latency via the network to a system that is highly dependent on latency as a key performance indicator seems unwise. However, Charter developed a system in which additional bandwidth can be used in conjunction with VR software techniques to help mask a lot of the additional latency. Motion-to-photon (MTP) latency is the delay between a motion and that motion being reflected in a frame shown on a display. MTP latency is one of the most important performance indicators of a VR system. Traditionally, a MTP latency of 20 ms or less is considered good. This means that when a user moves their head, having an updated frame within 20 ms is critical to ensure a smooth experience and avoid VR sickness. Ideally, a VR system is able to maintain a MTP latency that matches the desired framerate. In VR systems with 72 frames per second (FPS), the latency between each frame would be about 13.9 ms and about 11.1 ms for 90 FPS.

VR providers have developed software techniques such as reprojection to ensure the MTP latency is low even when a system cannot render at the target frame rate. Meta's (Oculus) implementation of this feature is called Asynchronous TimeWarp and Valve's (Steam) implementation is called Asynchronous Reprojection. The feature works by taking the latest motion data available and applying it to a rendered frame in order to reproject or warp the frame to a view that is closer to the user's current head location. The technique works because motion data from the headset's sensors are always available to the headsets with low latency (1-2ms). This latency is usually much lower than the render time, and in cases where the renderer is unable to hit the target framerate, reprojection allows a system to appear more responsive. Without this technique, users would always have to wait the entire render loop latency to see an updated view, but with it, the view can be updated more consistently. This technique has limitations including introducing artifacts in a frame, especially in cases where motion is high. Additionally, this technique only modifies the view; the actions of the game will still require the full render loop latency to display to a user. However, this limitation is well worth it since the main cause of VR sickness is not delays in viewing game actions (such as firing a gun), but instead delays in seeing an updated view after moving your head.

Providing a cloud VR service requires understanding the different latencies in the system and balancing tradeoffs between bandwidth, compute and latency. Figure 2 - Cloud Virtual Reality Latenciesshows the different latencies in a cloud VR system.

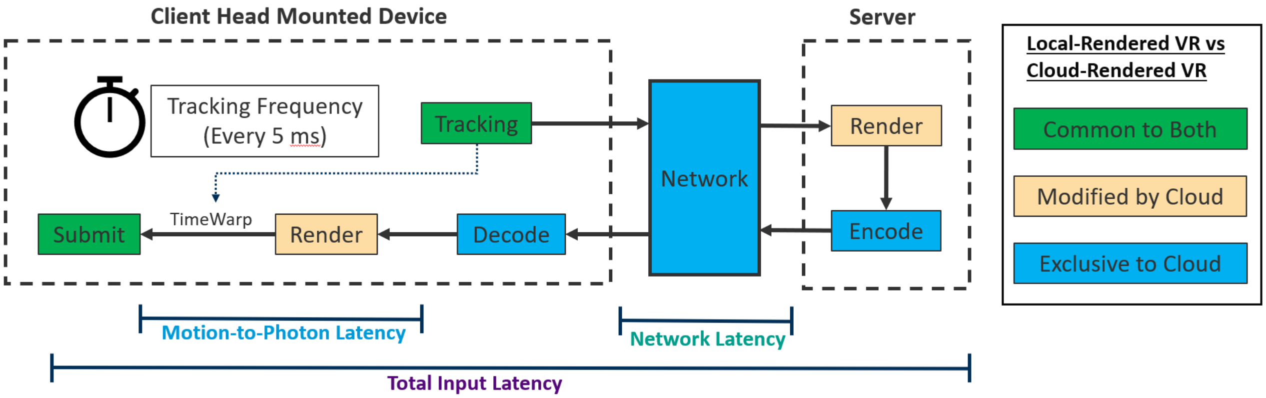

Figure 2 - Cloud Virtual Reality Latencies

The network latency is simply the round-trip time (RTT) between the client and server. The total input latency is the delay between any action (such as a button press) and when that action's result is reflected in the display. This is the main latency that suffers from cloud VR, however it is critical to separate this latency from the MTP latency. Traditional multiplayer gaming (PC or console) latencies range from 30-100 ms. Many gamers can tolerate latencies up to 100 ms depending on the type of game. The total input latency is most relatable to traditional multiplayer gaming latencies since those also include the system latency (PC or console) as well as the network latency. As previously mentioned, VR cannot tolerate latencies that high due to the nature of being head-mounted. This is why the distinction between total input latency and MTP latency is so important. MTP latency takes advantage of reprojection on the client to maintain a low perceived latency from head motion, regardless of total input latency.

Understanding these latencies allowed Charter to move forward with exploring how a cloud VR system could be delivered on our network. Beyond latency, delivering a 4K stream from a real-time rendered system requires a robust system and network capable of reliable high bandwidth. Not only is the resolution high, but the framerates used in VR (starting at 72 FPS) are higher than traditional video streaming. Netflix recommends at least 15 Mbps for 4K video streaming which is usually 24 FPS. This means a 4K VR stream could utilize approximately 50 Mbps. Additionally, real-time interactive content means the use of video techniques such as buffering is limited.

The cloud VR use case is a clear challenge for any network, requiring high bandwidth and low latency. This challenge is what motivated Charter to understand if our network is capable of handling this emerging use case or other use cases with similar requirements.

5. Cloud Virtual Reality Field Trial

Charter aimed to launch an employee field trial in 2023 in which select employees would be given a Meta Quest 2 headset that would stream VR games from a Charter RDC. The first step to achieving this was to develop the application and test it within a lab.

5.1. Field Trial Application Technology

The core technology of the application is relatively simple. The client application takes the position, orientation and button states of the controllers and headset and sends them over the network. The server runs VR games on a Windows operating system (OS) and encodes the rendered frames for transport back to the client. The client then decodes the frame for display.

Testing in the lab revealed the limitations and challenges of a system like this. Although reprojection enables a low MTP latency, it only works to a certain point. The more latency in the system, including network latency, the older the frame is that the client is reprojecting. This is dependent on the total latency and speed of the head motion. A frame that moved very little compared to the previous frame can tolerate more latency. If the difference between frames is too high, reprojection creates artifacts. This is because the reprojection only changes the view and cannot create new pixels. One example of an artifact is black bars on the side of the frame. For example, if a user is looking forward and suddenly rotates their head to the right, reprojection will move the view to the right and black pixels will be shown on the right edge of the frame. These black bars can become visible to the user and if bad enough, cause VR sickness due to lack of smooth motion.

Given this issue, Charter looked into how bandwidth could be used in conjunction with reprojection to minimize the latency of a cloud VR system. One technique used to help mitigate the artifact issue is called over-rendering. This is the concept of rendering a frame larger than what can actually be displayed. Combining a larger frame with reprojection allows the reprojection to have additional pixels that can be displayed to the user in the case of higher latency or higher motion. This was a key feature that enabled Charter to utilize more bandwidth to help reduce the perceived latency. While the Charter network was able to handle additional bandwidth, this bandwidth comes at the cost of higher compute. By rendering at a higher resolution, the server GPU requirements are increased.

Over-rendering was initially configured to render 20% more pixels, divided equally among all four sides of the frame. However, most of the time, additional pixels are only needed on one or two sides of the frame, depending on the direction of the head motion. To help alleviate the additional compute demands, another technique called head motion prediction (HMP) was developed. This technique utilizes a Kalman filtering algorithm to predict the direction of a user's head motion. In general, when a user moves their head to look in a direction, the movement will be in a single direction between most frames until they change directions. This provides a good opportunity for prediction in which errors are only expected in frames where the direction is changed. Combining prediction with over-rendering allows the server to over-render from a predicted head position. This enables a better use of rendered pixels, minimizing the percentage of over-rendering needed and lessening the compute requirements. This behavior is shown in Figure 3 - Over-rendering and Head Motion . The artifacts present are a function of head motion speed and network latency. The scenarios in the figure show the same scene with the same network latency and head motion speed and demonstrates how artifacts are reduced with over-rendering and prediction.

CloudVR
~40% Missing Frame

CloudVR with Over-Rendering
~25% Missing Frame

CloudVR with Head Motion Prediction
~0% Missing Frame

Figure 3 - Over-rendering and Head Motion Prediction

Utilizing these techniques in a field trial required a quantifiable way to measure artifacts and understand what quality of service can be delivered from the Charter network at different points. Charter created a turntable capable of rotating a headset at precise speeds in a lab environment where latency could be injected. This setup is shown in Figure 4 - Cloud VR Video Quality Measurement, below.

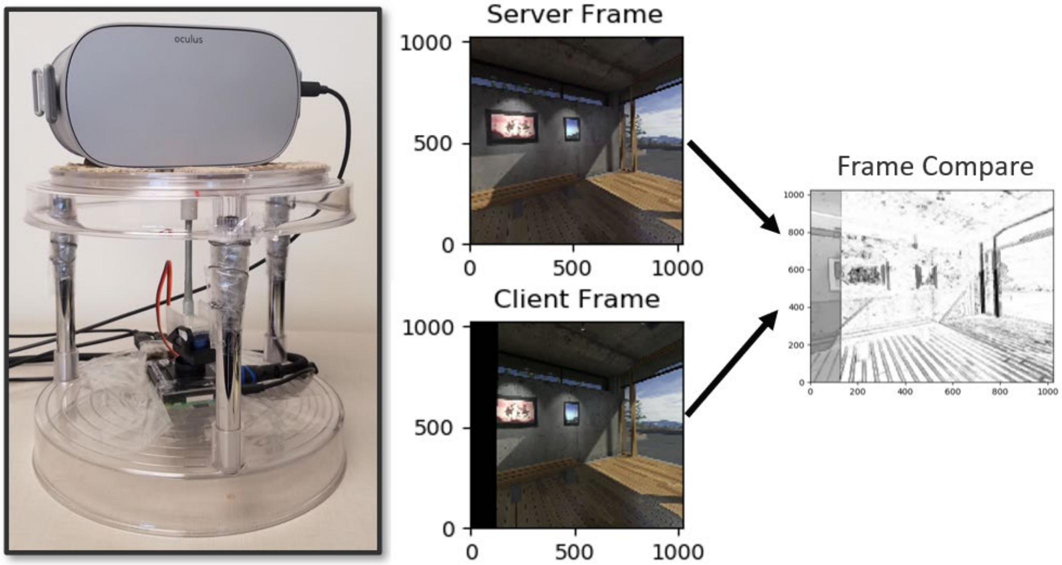

Figure 4 - Cloud VR Video Quality Measurement

This system included frame captures on both the client and server applications. A video quality metrics analysis was then done to automatically detect the artifacts introduced between the client and server. Aside from unavoidable frame differences caused by the transcoding, black pixels introduced from head motion would cause dramatic changes in quality scores. The lab testing provided key guidelines on the tolerable network latencies for different over-rendered resolutions at different head motion speeds. A game was considered high motion if it required head rotation of about 400 degrees per second. We found that at our target resolution, a network latency as high as 35 ms would still enable a good experience for most users.

5.2. Field Trial Infrastructure

A 35 ms target network latency enabled Charter to deploy servers in an RDC in which employees within a few hundred miles would achieve latencies well below the target. The field trial infrastructure included an Infrastructure as a Service (IaaS) made up of several server racks with data center-grade GPUs. The IaaS ran Windows 10 virtual machines (VM) which the content ran on. Connecting the client application to the server application utilized an orchestration layer that was developed for the field trial. Since these microservices were not latency sensitive, they were hosted in a Charter national data center (NDC) as part of Charter's Container as a Service (CaaS). The field trial architecture is shown in Figure 5, below.

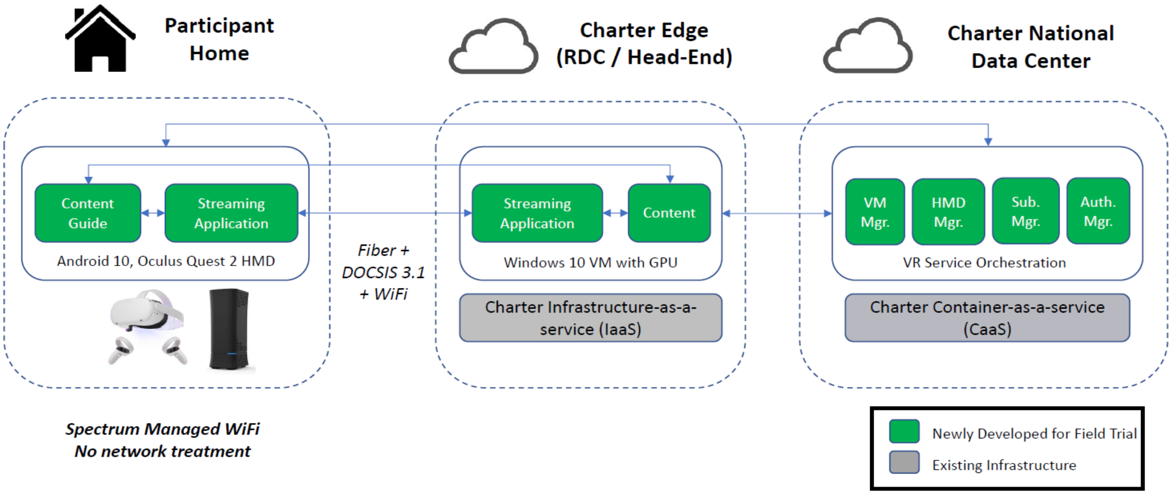

Figure 5 – Cloud Virtual Reality Field Trial

The client application utilized the orchestration layer to authenticate employees via the CaaS microservices and connect them to a VM hosted in the RDC. Once connected, the client is able to start streaming from the server to their device on the home WiFi.

5.3. Field Trial Results

The field trial included 35 employees over a 90-day period. The first phase of the field trial had no network treatment and the second phase had active queue management (AQM) activated for the users. The users combined for over 180 hours of data. Built into the application was a metrics gathering system that collected network and performance data of the application, home network and servers. After each session, the application prompted users to give a 1-5 star rating of the experience and content. The experience rating was meant to rate the quality of the stream while the content rating was meant to separate the user's opinion of the content. Users reported an average of 4.3 for the experience rating and 4.5 for the content rating.

5.3.1. Network Metrics

Sessions with lower network latency correlated to a higher experience rating. Low latency (<27 ms) sessions had 88% good experience ratings (4+) while high latency sessions (30+ ms) had 61% good experience ratings. Table 1 – Experience Rating vs Network Latency below shows the median network latency for each experience rating. Overall, the field trial averaged a network latency of 29 ms.

Table 1 – Experience Rating vs Network Latency

Rating	**5**	**4**	**3**	**2**	**1**
Median Network Latency (ms)	27.6	28.3	29.9	28.9	30

The average and median latency were lower in the second phase with AQM, but given a different set of users in each phase, it can't be determined if this was because of AQM conclusively. Additionally, network latency increased during peak times (7PM – 11PM) without AQM, but actually decreased during peak time when AQM was on.

Jitter was measured based on the client utilizing the method described in internet standard RFC 1889/3550. The average and median jitter was about 3.5 ms. Lab testing determined that jitter of 5 ms or less provided a good experience for most users. We found a correlation between jitter and network latency, in which higher network latency usually led to higher jitter.

We found that packet loss was arguably more important than network latency in terms of being a key performance indicator. Ideally, packet loss would be 0% at all times. However, given this field trial relied on home WiFi connections, some levels of packet loss were expected. Lab testing indicated an average of less than .1% of packet loss was ideal. The application did utilize forward error correction (FEC) to help mitigate higher levels (acceptable up to .5%). The field trial demonstrated a median of .03% packet loss, however we found that packet loss would happen in bursts. These bursts included periods of high packet loss (1%+) within the home which could cause visual impairments. These bursts were rare and, in most cases, happened in isolation within the user's home WiFi network. Users that experienced sustained levels of high packet loss were instructed to upgrade and/or replace their modem/router.

WiFi metrics were also gathered by the client since most internet issues in a home occur on the WiFi network. All users were instructed to utilize 5GHz WiFi, and the application would warn users that were connecting with 2.4GHz. However, since this was only a warning, some users proceeded on 2.4GHz regardless and demonstrated higher network latency, jitter and packet loss. The WiFi signal strength was measured with the network latency, which were found to be correlated. Sessions with higher signal strength (closer to the router) were found to have lower network latencies.

The target streaming bitrate was 50 Mbps and averaged about 45 Mbps during the field trial. The actual bandwidth used is determined by the framerate that could be achieved by the server and was dependent on the rendering performance of each game. The upstream bandwidth used on average was 600 Kbps, higher than required for a system like this due to the metrics being uploaded from the client.

5.3.2. Client and Server Performance

The average total input latency was 74 ms and 88% of sessions performed under 80 ms. Other than network latency, the largest latencies were the render and encode latencies on the server. Optimizations and improvements to the application can reduce the total input latency further, which would allow for a larger tolerance for higher network latency.

GPU utilization averaged less than 50% however the game with the highest fidelity graphics averaged 65%. Encode performance was arguably the bottleneck of server performance in the field trial, with encode utilization averaging 53%. Encode performance was also dependent on content, where games with higher motion had longer encode times.

Overall, head motion prediction performed well. Only 7% of frames reported a prediction error, in which the predicted render origin was not close enough to the actual origin to prevent any black pixel artifacts. Prediction in the Y direction (up and down) was more accurate than in the X direction (left and right). This is likely because left to right head movement is more common than up and down head movement in VR. Prediction performance was less accurate for content with higher motion.

5.3.3. Survey Results

Employees were surveyed after the field trial to gather feedback on the service. Twenty-five out of 27 respondents said they were satisfied or more than satisfied with the performance; 22 out of 27 users said they would likely subscribe to a service like this; 17 out of 22 users said the service was consistent or very consistent; and 13 out of 18 users (with previous VR experience) said this was equal to or better than previous VR experiences.

6. Conclusion

Charter aimed to test the network against a demanding use case that utilized very high bandwidth and required consistent performance with low latency. The results showed that Charter's network is already capable of providing quality Cloud Virtual Reality service from a Regional Data Center to users nearby and software improvements and reductions to access network latency will expand the service to more users.

The physical range between the server and users remains a constraint that can be further reduced via software improvements and optimizations, more efficient future hardware and reduction of network latencies. More efficient hardware is critical for making services like this more economical. Currently, the upfront cost of the hardware (especially GPUs) makes a service like this a challenge to scale.

Network key takeaways included the importance of the in-home WiFi network, ensuring users are near their router and that packet loss in the home, caused by congestion, is minimized as much as possible. Sessions with lower network latency and jitter resulted in higher ratings. More so than bandwidth, this field trial demonstrated the importance of understanding network latencies and reducing them as much as possible.

Services like cloud VR that demand low latency stand to benefit from new network technology, such as LLD and L4S. Cloud VR technology will continue to be improved and be tested against the latest network technologies.

Abbreviations

VR	virtual reality
AR	augmented reality
RDC	regional data center
NDC	national data center
RSSI	received signal strength indicator
GPU	graphics processing unit
CPU	central processing unit
LLD	low latency DOCSIS
DOCSIS	data over cable service interface specification
AQM	active queue management
L4S	low latency, low loss, and scalable throughput
FOV	field of view
PC	personal computer
MTP	motion-to-photon
FPS	frames per second
RTT	round-trip time
OS	operating system
HMP	head motion prediction
IaaS	infrastructure as a service
CaaS	container as a service
VM	virtual machine
AQM	active queue management
FEC	forward error correction

Bibliography & References

"Internet connection speed recommendations," Netflix, https://help.netflix.com/en/node/306

"Under the hood: Building 360 video," Engineering at Meta, Oct. 15, 2015. https://engineering.fb.com/2015/10/15/video-engineering/under-the-hood-building-360-video/

"How 10G Networks Will Power Holographic Technologies | NCTA — The Internet & Television Association," www.ncta.com, Dec. 08, 2020. https://www.ncta.com/whats-new/how-10g-networks-will-power-holographic-technologies

"Keeping the virtual world stable in VR | Qualcomm," www.qualcomm.com, Jun. 28, 2016. https://www.qualcomm.com/news/onq/2016/06/keeping-virtual-world-stable-vr

"Asynchronous TimeWarp (ATW) | Oculus Developers," developer.oculus.com. https://developer.oculus.com/documentation/native/android/mobile-timewarp-overview/

"Introducing SteamVR Motion Smoothing," Valve Corporation, Nov. 27, 2018. https://steamcommunity.com/games/250820/announcements/detail/1705071932992003492

H. Schulzrinne, S. Casner, R. Frederick, and V. Jacobson, "RFC 3550 - RTP: A Transport Protocol for Real-Time Applications," Internet Engineering Task Force, Jul. 2023. https://www.ietf.org/rfc/rfc3550.txt