The Effects of Latency on Player Performance in Cloud-based Games

Mark Claypool and David Finkel
Computer Science and Interactive Media & Game Development
Worcester Polytechnic Institute
Worcester, MA, USA
Email: {claypool,dfinkel}@cs.wpi.edu

Abstract—Cloud-based games are an increasingly popular method to distribute and play computer games on the Internet. While there has been some work studying network aspects of cloud-based games and examining the effects of latency on traditional games, there has not been sufficient research on the impact of latency on cloud-based games nor a comparison of the impact of latency on cloud-based games versus traditional games. This paper presents the results of two user studies that measure the objective and subjective effects of latency on cloud-based games, one study using the commercial cloud game system OnLive and the other study using the academic cloud game system GamingAnywhere. Analysis of the results shows both quality of experience and user performance degrade linearly with an increase in latency. More significantly, latency affects cloud-based games in a manner most similar to that of traditional first-person avatar games, the most sensitive class of games, despite the fact that the cloud-based games may have a different user perspective. These results have implications for cloud-based game designers and cloud system developers.

I. INTRODUCTION

The increase in power and connectivity of computer networks coupled with the growth in popularity of computer games has enabled cloud-based games as a new method for distributing and playing computer games. Instead of distributing games through traditional media such as DVDs or electronic download, cloud-based games keep the game content itself on servers, sending down the game images as video to the client where the user reacts, then sending up the user input to the servers. These cloud-based systems give game providers more control over the content, while allowing users access to a large game library using only a single, lightweight client.

Cloud-based games require video-type network capacities down to the client, similar to that of streaming video, and much more modest capacities up from the client [1], capacities that are available to many game clients. Moreover, cloud-based games are relatively robust in the face of modest packet loss [1]. This leaves latency as the dominant network degradation that can affect a user’s experience playing cloud-based games.

There have been numerous studies on the effects of latency and games (e.g., [2], [3], [4]). In addition, there have been at least as many studies on the effects of latency on interactive, multimedia applications such as audio and video conferences. Yet, while cloud-based games are similar to traditional games in terms of user interactions and are similar in downstream bitrate characteristics to videoconferences, there are fundamental differences in the user’s interaction with the media compared to either videoconferences or traditional games. Specifically, the effects of latency on video conferences are often determined by the impact of latency on the spoken dialog that accompanies the video. Such voice-based dialog interaction is not a major component of most computer games. The effects of latency on traditional games depends upon the genre of the game, with direct game interactions typical of games with a first-person perspective being most sensitive to latency, while games with a third-person or omnipresent perspective being less sensitive to latency [5]. Yet the lightweight client in cloud-based games cannot respond to user input in the same fashion as can a client in a traditional game, so the impact on latency based on game genre may be different. To the best of our knowledge, there has yet to be a scientific exploration of the effects of latency on cloud-based games, specifically comparing them to traditional games.

This paper presents details on two user studies that measure the impact of latency on cloud-based games. One study measured the effects of latency using a commercial cloud-based game system, OnLive, while the other used an academic cloud-based game system, GamingAnywhere. Both studies had over 30 users play short game sessions with a third-person avatar game, with the users blind to controlled amounts of added latency. During each session, the system measured and recorded user in-game performance, and after each session users provided subjective opinions on the gameplay.

Analysis of the results shows player performance degrades directly with an increase in network latency. Applying a linear regression to the data finds that every 100 milliseconds of latency results in about a 25% decrease in player performance. While this result in itself is perhaps not surprising, what is interesting is the degree with which latency affects the player experience. In particular, latency affects player experience with cloud-based games most similarly to that of traditional first person perspective games, even though the games tested have a third person perspective. This has implications for system designers as traditional third person perspective games, perhaps the most common genre of computer game, can generally tolerate up to 500 milliseconds of latency, while traditional first person perspective games are much more sensitive to latency, often degrading player performance even with as little as 100 milliseconds of latency.
The rest of this paper is organized as follows: Section II presents work related to this research; Section III describes our two user studies that measure the impact of latency on cloud-based games; Section IV analyzes the results of the experiments; and Section V summarizes our conclusions and presents possible future work.

II. RELATED WORK

Previous research related to the work in this paper broadly falls into three areas: latency and traditional games, cloud-based games and latency and cloud-based games.

The effects of latency on traditional games has been studied for many game genres, including car racing [2], role playing games [3] and first person shooters [4]. While such work has helped better understand the impact of latency on traditional games, the results may not hold for cloud-based games that have a fundamentally different underlying system architecture. Still, results from such work is used for comparison of the effects of latency on cloud-based games in this paper.

Shea et al. [6] conducted a systematic analysis of cloud-based game platforms and specifically measured responsiveness in OnLive with added latency. While the authors found system delay rose linearly with latency, they did not measure the impact of delay on users playing cloud-based games.

Chen et al. [7] discussed the effects of network latency, packet loss, and bitrate on frame rates and graphic quality for cloud-based game systems OnLive and StreamMyGame. While the authors found frame rate and graphics correlated with user quality of experience, they did not explicitly measure player performance with added latency.

Jarschel et al. [8] conducted a user study in an emulated cloud gaming system, measuring the quality of experience for games users selected to play. While closely related to the work in this paper, the authors did not compare cloud-based games with latency to traditional games.

III. APPROACH

In order to measure the effects of latency on player performance, we conducted two separate user studies, presented in Section III-A and Section III-B.

A. OnLive

The first study used a commercial cloud-based game system called OnLive.1 With OnLive, users install the lightweight client application on their PCs (or purchase a lightweight game console) and connect to the OnLive cloud. Once a player selects a game to play, the gameplay is rendered by a server in the cloud and streamed down to the user’s computer, with user interactions with the game being transmitted up to the server where it is incorporated into the game.

Our OnLive user study had users play the 1999 arcade hit called Crazy Taxi.2 Figure 1 depicts the user’s view of the taxi in Crazy Taxi. The goal of Crazy Taxi is to earn points by driving a taxi cab and delivering as many customers to their chosen destinations as quickly as possible. Along the way, users can earn extra points by performing stunts such as near misses with other vehicles or jumping over ramps. The taxi cab is controlled via the OnLive game pad. The objective measure of performance is the points the user earns during the allotted time. We set the game options to have the user start in the same position each time and with a traffic difficulty so that each game took about 1 minute, based on pilot studies.

Users played Crazy Taxi on the OnLive game console displayed on a 65” LCD TV. The console was networked to a laptop, configured to route traffic to the University LAN and then to the OnLive cloud via the Internet.

The laptop was a Macbook pro running OS X 10.9.2 with an Intel i7 2.3 GHz processor, 4 GB of RAM, a Broadcom 57765 Gbit Ethernet card and a UTechSmart USB 2.0 100 Mbit Ethernet adaptor. The laptop was configured for Internet sharing, allowing the console to connect through the Ethernet card and forward traffic via the USB Ethernet adaptator. The University has a 10 Gbit/s LAN backbone with about a 1 Gbit/s connection to the rest of the Internet.

Latency between the console and the OnLive cloud was controlled by Dummmynet [9] running on the laptop, automated with a script to control latency based on the trial schedule for the user study. The baseline latency to the OnLive servers was about 50 milliseconds. Based on pilot studies, seven different latencies were chosen, from 0 to 150 milliseconds in steps of 25 milliseconds, with the added latency split evenly between the uplink and downlink. Values of 0, 75 and 150 milliseconds were repeated to provide 10 trials total. The order of induced latencies were manually shuffled: 0, 50, 25, 75, 150, 100, 125, 75, 0, and 150 milliseconds.

Upon arriving for the study, users were read a short instructional script and given an iPad for providing demographic information via a Web-based survey constructed using Qualtrics.4 Demographic information included gender, age and major as well as previous experience playing computer games.

Users were instructed to play a practice round of Crazy Taxi, 5 minutes of normal game play without added latency. After the practice round, the latency trials commenced, each taking about 1 minute. The iPad was used to gather subjective feedback after each trial.

1http://www.onlive.com/
2http://en.wikipedia.org/wiki/Crazy_Taxi
4http://www.qualtrics.com/
Users were solicited from the University campus through a mass email and by advertising posters. In addition, participants were recruited from the Psychology department’s Participant Pool, a system whereby Psychology students participate in user studies to obtain class and major credit. For added incentives, participants were provided with refreshments during the study and had a chance to win a raffle for a Visa gift card worth $50. In total, 49 users participated in the OnLive study.

B. GamingAnywhere

The second study used an academic cloud-based game system called GamingAnywhere. With GamingAnywhere, users install the server and client software along with the game to be played. The system then plays the game on the server, sending the rendered frames down to the client. At the client, the user’s game commands are captured and sent back up to the server where the commands are fed into the game as if they came from a local user.

Our GamingAnywhere user study had users play the open source marble-roller called Neverball. Figure 2 depicts the user’s view of the ball in Neverball. The goal of Neverball is to tilt the game world to roll a marble around obstacles and to a marked goal. The only controls are the arrow keys, making it relatively easy for even novices to understand the controls and play the game. The objective measure of performance is the time to get the ball to the goal.

Based on pilot studies, we chose level seven from the “easy” Neverball level set. The level, depicted in Figure 3, is composed of an undulating ramp with maze-like fences on the surface.

GamingAnywhere provides documentation on how to construct client and server configuration files, as well as provides pre-made configuration files for several games. We used the Neverball configuration files provided by GamingAnywhere for our user study.

The server ran on a Microsoft Windows laptop PC with an Intel i7 2.4 GHz processor, an Nvidia GeForce GTX 675M graphics card and 16 GB of RAM. Pilot studies indicated the server could render and stream Neverball at 50 frames/s at 800x600 pixels of resolution, the resolution used in our study.

The client ran on a Microsoft Windows desktop PC with an Intel i7 2.4 GHz processor, Intel onboard graphics card and 8 GB of RAM. The client could smoothly receive and display frames at all rates tested by the server.

The client and server were connected by a 1 Gbit/s Ethernet crossover cable. Latency between the client and server was controlled with Dummynet running on the server, automated with a short script that set up the appropriate Dummynet rules and pipes during each user session. The baseline latency from the client to the server is less than 1 millisecond. Based on pilot studies, six different added latencies were chosen: 0, 33, 66, 100, 150, and 200 milliseconds, with the added latency split evenly between the up and downlink.

At the beginning of each session, each user was read a script providing instructions for the game and study. Users were then prompted to provide demographic information including their age, gender, and previous game experience.

Users were given an initial practice round at 0 ms of latency to learn to play the game. Subsequent latencies were based on player groups. The latencies were divided into three groups: 1) 33 ms and 66 ms, 2) 100 ms, and 3) 150 ms and 200 ms. Each user played the level with each network latency in order of one of the six combinations of these groups. For instance, if one user played the trials in the order 33 ms, 66 ms, 100 ms, 150 ms, and 200 ms, the next participant would play the trials in the order 33 ms, 66 ms, 150 ms, 200 ms, and 100 ms.

For each trial, the user’s time to complete the level was recorded using Neverball’s in-game timer along with subjective questions about the trial.

The total time for a user to complete the user study session was about 10 minutes.

Users were solicited from the University campus through a mass email and by word of mouth. As an added incentive, participants were told they would have a chance to win a raffle for a Newegg.com gift card worth $75. In total, 34 users participated in the GamingAnywhere study.
C. Summary

In summary, we conducted two user studies on cloud-based game systems with games that were easy to learn, but not likely to be frequently played by participants. One study used OnLive with a console connecting to the OnLive server in the cloud, and the other study used GamingAnywhere with a client and server on a private LAN. Both studies induced latency between the client and server via Dummynet with latencies ranging from 0 to 200 milliseconds. User in-game performance at various latencies provided objective measures, and user opinions on the gameplay provided subjective measures. Users were solicited via advertising, email and some modest incentives, with over 30 users participating in each study. Table I provides a concise summary of the two user studies.

<table>
<thead>
<tr>
<th>Cloud System</th>
<th>OnLive</th>
<th>GamingAnywhere</th>
</tr>
</thead>
<tbody>
<tr>
<td>Game Used</td>
<td>Crazy Taxi</td>
<td>Neverball</td>
</tr>
<tr>
<td>Hardware</td>
<td>Console, TV, laptop</td>
<td>PCs</td>
</tr>
<tr>
<td>Latency Control</td>
<td>Dummynet</td>
<td>Dummynet</td>
</tr>
<tr>
<td>Latencies (msec)</td>
<td>0, 25, 75, 100, 125, 150</td>
<td>0, 33, 66, 100, 150, 200</td>
</tr>
<tr>
<td>Number of Users</td>
<td>49</td>
<td>34</td>
</tr>
</tbody>
</table>

Table I: Summary of Setup for User Studies

Note, the latency values, and to a lesser extent the subjective questions, were different for each game since: 1) two different, independent research teams conducted the experiments, and 2) the latencies were selected based upon pilot studies with different games (i.e., Neverball vs. Crazy Taxi) and different underlying systems (i.e., GamingAnywhere vs. OnLive).

IV. Analysis

This section presents the analysis of the GamingAnywhere and OnLive user studies that can be compared with each other. Specifically, this section provides analysis of user demographics (Section IV-A), and subjective (Section IV-B) and objective results (Section IV-C). Additional in-depth analysis specific to each study can be found in the appropriate technical reports [10], [11].

A. Demographics

For the OnLive Crazy Taxi study, about 95% of the users were 18-22 years old and about 70% were male. About 75% self-reported an “average” or more experience playing computer games.

For the GamingAnywhere Neverball study, all participants were 18-22 years old and about 90% were male. All self-reported an “average” or more experience playing computer games.

B. Subjective

For the OnLive Crazy Taxi study, user subjective opinions were provided on a 7 point scale (1 - worst, 7 - best) and combined into a quality of experience metric (QoE) as in [12]. Figure 4 depicts the results. The x-axis is the added latency in milliseconds and the y-axis is the QoE. The points are the QoEs at the indicated latencies, shown with standard error bars. The connected points look almost linear, sloping down left to right and, in fact, a linear regression has an $R^2$ of about 0.92.

For the GamingAnywhere Neverball study, user subjective opinions were provided on a 5 point scale (1 - worst, 5 - best) with an explicit question about quality of experience. The Friedman test was applied to the results, using a null hypothesis that the means for every trial are equal (i.e., that there is no correlation between latency and the mean of the survey responses) – the alternative hypothesis represents the condition that the means are different. The test was statistically significant at p=0.002, indicating latency affected user QoE. Figure 5 depicts the results. The x-axis is the added latency, in milliseconds, and the y-axis is the average of the rank order for the QoE for that latency. The points are the QoEs at the indicated latencies, shown with standard error bars. The connected points look somewhat linear with a $R^2$ of about 0.86, with a more significant drop between 100 and 150 milliseconds.

C. Objective

For the OnLive Crazy Taxi study, the objective measure of performance was points earned delivering taxi customers to their destinations. The higher the points, the better. Figure 6 depicts the results. The x-axis is the added latency in milliseconds and the y-axis is the user points. The data points are

---

8 Although the OnLive Crazy Taxi study used a 7 point scale, the specific QoE values are not directly compared with each other.

the scores averaged across all users at the indicated latencies, shown with standard error bars. As for QoE, the connected points look almost linear, albeit with a slightly steeper slope, with an $R^2$ of about 0.87.

For the GamingAnywhere Neverball study, the objective measure of performance was time to roll the marble to the goal. Note, for Neverball, the lower the time, the better. Figure 7 depicts the results. The x-axis is the added latency in milliseconds and the y-axis is the time to get the marble to the goal in seconds. The data points are the times averaged across all users at the indicated latencies, shown with standard error bars. As for earlier graphs, the connected points look almost linear, albeit sloping upward, with a slightly flatter shape on the right side. The $R^2$ is a more modest 0.70.

In order to better understand the impact of latency on players in cloud-based games, it is helpful to compare the user study results to user study results from traditional network games. Previous work has found the effects of latency depend upon user in-game perspective [5]. The game perspective defines how a user views the game world on a screen.

With an avatar-interaction perspective, the user interacts with the game through a single representative character, called the avatar. Games with an avatar-interaction typically have either a first person perspective where the user sees the game world through the eyes of the avatar, or a third person perspective where the user follows an avatar in the virtual world. First person shooter (FPS) games, role-playing games (RPGs), sports games and racing games are all examples of game genres that have an avatar-interaction perspective. These game genres often differ in the perspective – for example, FPS games have a first person perspective while RPGs typically have a third person perspective.

With an omnipresent perspective, the user has the ability to view and interact with different aspects of the game world. The user is said to be omnipresent in that his/her actions have a more global influence than actions in an avatar model. The perspective of games with the omnipresent interaction model is often variable, giving users an aerial perspective to provide a bird’s eye view of the virtual world, but also allowing users to zoom in to a third person perspective to provide finer granularity of control over individual resources. Real-time strategy games (RTS) and construction and simulation games are examples of game genres with the omnipresent perspective.

In order to compare the effects of latency on cloud-based games to traditional games as well as compare the user performance with latency for Crazy Taxi and Neverball, user objective performance results are normalized from 0 (worst) to 1 (best). Results from previous studies of latency and traditional network games (first person avatar [13], [2], third person avatar [3], [14], and omnipresent [15]) are similarly normalized and fitted with an exponential curve [5].

Figure 8 depicts the results, summarizing classes of traditional network games. The horizontal gray rectangle is a visual indicator of user tolerance for latency. Gameplay quality is generally acceptable above the gray area and unacceptable below it. The exact latency tolerance threshold depends on the game and to some extent the users own perception and sense of immersion (hence the gray color and the rectangle shape rather than a line). The data points are the normalized objective performance measures for Crazy Taxi (squares) and Neverball (pluses).

For traditional networked games, games with the avatar model of user interaction are more sensitive to latency than games with the omnipresent model. Deeper within the avatar model, games with a first person perspective are more sensitive to latency than games with a third person perspective.

---

10Providing details on these studies is not feasible given the space constraints, but the interested reader is encouraged to follow the references.
For cloud-based games, both Crazy Taxi and Neverball most closely follow the latency degradation curve of traditional first person avatar games, despite both games having a third person perspective.

We surmise this result stems from the nature of user interaction with cloud-based games that differs from that of traditional games. In a traditional game, a user action can be handled immediately by the local game engine, and the indirect nature of a third person game makes the effects of latency less severe than for a first person game. However, in a cloud-based game, user actions cannot be shown until the action is transmitted up to the cloud, acted upon and rendered by the server, and transmitted back to the client. Thus, the responsiveness of a cloud-based game is directly impacted by even small amounts of latency, similar to the high latency sensitivities of traditional first person games, even when playing a third person or omnipresent game.

V. CONCLUSION

Cloud-based game systems have become increasingly relevant given the growth in network connectivity and interest in convenient game distribution models for game content providers and game players. While the effects of latency on traditional network games is fairly well studied, the impact of latency on cloud-based games is less well understood, particularly when comparing latency impact to both game types.

This paper presents the results of two user studies that explore the impact of latency on cloud-based games. The two studies use two popular cloud-based game systems: OnLive, a commercial cloud-based game system, and GamingAnywhere, an academic cloud-based game system – both used in a setup that allows for controlled amounts of added latency. Two third person avatar games are studied, Neverball and Crazy Taxi, providing measures of user performance and opinions on quality of experience. Careful analysis of the results provides two main contributions of this work: 1) specific results of the objective and subjective impact of latency on cloud-based game systems currently in use; and 2) comparison of the impact of latency on user performance in cloud-based games to the impact of latency on user performance in traditional network games.

The results show cloud-based games are sensitive to even modest amounts of latency, with user performance degrading by up to 25% with each 100 milliseconds of latency. Analysis comparing cloud-based games to traditional games shows cloud-based games are as sensitive to latency as first person avatar games, the most sensitive class of games to latency, even though the user perspective would suggest more tolerance for latency based on results for traditional network games. These results have implications for cloud-based game providers as they must be aware of system latency for all games, even those that have traditionally been more forgiving to network latencies such as third person or omnipresent games.

For future work, additional user studies with cloud-based games from other genres would help flesh out the results in Figure 8. Player expectations for arcade-style games (such as those used in this study) may differ from those of traditional network games, and additional studies could explore how these expectations might impact QoE for cloud-based games. The sensitivity of cloud-based games to latency suggests applying latency compensation techniques to mitigate the effects of latency on the player. However, most latency compensation techniques have been designed and tested for traditional networked games, so new research may be needed to study, and perhaps develop, new techniques to compensate for latency in cloud systems. Study of the effects of variation in latency (i.e., jitter) on cloud-based games is also possible future work.

ACKNOWLEDGMENTS

We would like to recognize and thank the students that did the work designing and conducting the user studies as well as providing the preliminary analysis: James Anouna, Zachary Estep and Michael French for the GamingAnywhere Neverball user study and Robert Dabrowski, Chrisitan Manuel and Robert Smieja for the OnLive Crazy Taxi user study.

REFERENCES