Latency Thresholds for Usability in Games: A Survey

Kjetil Raaen
Westerdals Oslo School of Art, Communication and Technology
Faculty of Technology
University of Oslo, Institute for Informatics
Simula Research Laboratory

Tor-Morten Grønli
Westerdals Oslo School of Art, Communication and Technology
Faculty of Technology

Abstract
User interactions in interactive applications are time critical operations; late response will degrade the experience. Sensitivity to delay does however vary greatly with between games. This paper surveys existing literature on the specifics of this limitation. We find a classification where games are grouped with others of roughly the same requirements. In addition we find some numbers on how long latency is acceptable. These numbers are however inconsistent between studies, indicating inconsistent methodology or insufficient classification of games and interactions. To improve classification, we suggest some changes.

In general, research is too sparse to draw any strong or statistically significant conclusions. In some of the most time critical games, latency seems to degrade the experience at about 50 ms.

1 Introduction
Whenever a human interacts with a computer, the computer could be said to run an "interactive application". A user enters some input and the computer responds. Word processors, spreadsheets and web-browsers are based on a workflow where the user continuously enters input, and the system responds immediately. Conversely, not all applications have this interaction as its central function. Simulations running on supercomputers spend very little time interacting with the user, and most of its time is spent doing calculations; these applications are often said to do "batch processing".

Games are a class of applications that commonly requires significant amounts of relatively fast interactions. They are also among the few applications where users regularly measure and worry about latency since this has a high impact on gameplay.

This paper was presented at the NIK-2014 conference; see http://www.nik.no/.
This makes games one of the most interesting categories of applications for studies on delays.

Research has often focused on response time as a critical factor for quality of experience (QoE) in networked games. This metric is equally relevant for local games, but achieving sufficient response time is rarely a problem in this class of applications.

The dominating factor of response time in networked games and interactive worlds are usually network latency; hence earlier work is mostly focused on network metrics. Conversely in local games, lacking any communication, response time is only bound by computation, which is a known factor when creating the game and can thus be compensated for.

Data on acceptable latency in games, or on how player performance correlates with latency is sparse in current literature. In some cases multiple papers cite the same few sources of empirical data.

Work on improving latency of games often quote 100ms as acceptable latency for fast-paced games, for example Raaen et al. 2012 [19]. Other types of games are quoted with more lax latency requirements. Our goal is finding if this is reasonable, or if there are other, better estimates. Games are very different in how fast reactions are necessary to play the game effectively. It is reasonable to assume that the response time they require depends on the speed of gameplay. The primary goal of this paper is to investigate the field of latency in games through a survey of existing literature. This will contribute to defining latency thresholds from a theoretical perspective and can further be used in research projects aiming to improve interactive applications.

2 Background

In any interactive application, there will be a delay between the time the user sends input and the result appearing on screen. This delay is commonly termed "input lag". If the computer communicates through a network, any communication over this network takes time, called "network latency" or simply "latency". Most players typically use the term "lag" for both types of delay. This can make complaints from users somewhat ambiguous.

Types of Games

"Latency and player interaction in online games" [8] by Mark Claypool and Kajal Claypool 2006), describes a whole set of game types and situations and recommend latency limits, dividing games into three broad categories:

**First Person Avatar** Describes games where the player controls an avatar, and the game is displayed from the point of view of that avatar, as if the player sees "through the eyes" of the avatar. The most common variant of this class of game is the **First Person Shooter** (FPS).

**Third Person Avatar** These games are also based on a single avatar controlled by the player, but the avatar is seen from the outside. The most studied variant in this group is the **Massive Multiplayer Online** (MMO) game.

**Omnipresent** In these games, the player does not control a specific avatar, rather multiple elements in a large game world. The most popular variant of this is the **Real Time Strategy** (RTS) game.
It is assumed that each of these classes of games has different latency requirements. The MMO genre is a particularly interesting case, and the games themselves as well as the technology behind them have received significant attention from computer scientists.

One of the main attractions of these games lies in the number of players that participate in the game. The more players that are in the game world, the more interactive, complex and attractive the game environment will become. [22]

Though MMOs are technically the most technologically demanding, Claypool does not place them in the most latency sensitive category.

Cloud Gaming vs. Client-Server

Client-server games are all designed from the ground up to handle network latency, deviation or displacement in phase timing and jitter (understood as irregular variation). By employing various prediction techniques and allowing a looser consistency in state between different players, these games will be able to alleviate, or in some cases completely isolate, the players from the effects of latency. Some of the most ubiquitous techniques are described in Bernier [4].

Initially, this paper describes the basic client-server case, where the client transfers data to the server; the server does necessary processing and sends back the result, which the client renders. However, strict adherence to the client-server pattern is rare except in purely experimental games. By allowing the client to do some calculations locally, feedback can be much quicker. Usually referred to as Client Side prediction, this solution will, at its simplest, run exactly the same code on the client as would run on the server. Whenever an update is received from the server, client state is reverted to conform to this data. Furthermore, instant effects, such as weapons firing, are executed immediately on the local client to give the player a feeling of responsiveness. To allow players to hit where they aim, the system needs to predict where other players are at a given moment, increasing complexity of the system even further.

The emerging concept of cloud gaming represents "Software as a Service" [21] where the software is a game. Delivering games this way presents some unique challenges, possibly including different requirements on the network latency. Latency in a cloud gaming scenario will work differently than latency in a traditional client-server game. Here, the client more closely resembles a traditional “thin client”, which simply forwards commands from the user to the server, and displays video from the server. The cloud, or remote server, performs all game logic and rendering tasks. In this scenario, requirements on client hardware will be very low. As tradeoff, the requirements on the network link are significantly stricter. There are at least two clear reasons for this. First, transferring high definition video requires significantly higher throughput than simple control signals; secondly, none of the techniques to alleviate network delay described above can be applied. Evaluations on these requirements are scarce at the moment.
3 Findings

The goals of this study is to investigate the issue of latency in games based on previous work and identify potential avenues for further empirical research on this topic. In order to do this we surveyed the area, searching IEEE, ACM and Springer databases, and selected papers were the title, keywords and abstract implied that the research focused on acceptable latency limits. Further, we traced the actual data presented in each paper to its original sources and included this paper. Filtering of relevant papers were based on the criterion on empirical data about how players react to latency in games. Table 1 list main contributors to the area. These are further presented in the rest of this section.

Controlled Studies

The oldest paper in this survey [12] dates back to 2001. Running a FPS game-server open to the Internet, the authors observe player behaviour and how this is affected by network delay. Most players connecting to their server fall into the interval 50 ms to 300 ms. Beyond this their only result relevant here is that players with delays over 400 ms seem much more likely to leave immediately.

Panthel and Wolf [17] study racing games, claiming this to be the most sensitive class of game. Using a setup with two identical machines, with controlled latency between them, the authors run two very different experiments. First they set up an identical starting position and perform identical actions on both sides, observing discrepancies. In the games studied, this leads to both players seeing themselves in the lead at the same time, even at the lowest tested latency of 100 ms. Next, they allow players to play actual games under varying latencies to the server. They find that the average player’s performance deteriorates first, at their lowest tested latency of 50 ms. The beginners drive too slowly to notice this delay, while excellent players are able to compensate for more latency. Performance of the excellent drivers degrade sharply at 150 ms.

Perhaps the most cited paper in this study is "Latency and player interaction in online games" [8] by Claypool and Claypool. Most authors citing this paper simply uses it as an explanation for setting acceptable limits to response times from the system they are evaluating. The value 100ms is frequently quoted. Among these are [6] which we discuss elsewhere in the literature review.

However, Claypool’s work and conclusions are much more nuanced, categorising games based on how they interact with the player. The paper describes a whole set of game types and situations and recommend latency limits. These limits are estimated at 100ms for "first person avatar" games, 500ms for “third person avatar” games, and 1000ms for "omnipresent" games. Further, the authors analyse different actions within each type of game. For each of these items, they have used empirical studies showing how player performance varies with network latency. The commonly quoted 100ms figure is based on multiple types of actions in a first person avatar game, the most latency-sensitive class of games according to this paper. It is important to notice that 100ms represents a point where player performance already has dropped off sharply from the previous data point 75ms. Because systems should be designed to withstand worst-case scenarios, the design goal should be below 75 ms. This paper is however mostly a secondary source, citing data from earlier work.

Though the 100ms estimate for "first person avatar” games is usually given with reference to [8], that study actually cites results from a paper by Beigbeder et al.
Table 1: Papers included in this study, the type of game studied and type of study.

<table>
<thead>
<tr>
<th>Authors/Year</th>
<th>Title</th>
<th>Game Type</th>
<th>Study Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>Henderson 2001 [12]</td>
<td>Latency and user behaviour on a multiplayer game server.</td>
<td>FPS</td>
<td>semi-controlled experiment</td>
</tr>
<tr>
<td>Nicholas and Claypool 2004 [16]</td>
<td>The Effects of Latency on Online Madden NFL Football.</td>
<td>Sports game²</td>
<td>controlled experiment</td>
</tr>
<tr>
<td>Quax et al. 2004 [18]</td>
<td>Objective and subjective evaluation of the influence of small amounts of delay and jitter on a recent first person shooter game.</td>
<td>FPS</td>
<td>controlled experiment</td>
</tr>
<tr>
<td>Claypool and Claypool 2010 [9]</td>
<td>Latency Can Kill: Precision and Deadline in Online Games.</td>
<td>various</td>
<td>controlled experiment</td>
</tr>
<tr>
<td>Amin et al. 2013 [1]</td>
<td>Assessing the impact of latency and jitter on the perceived quality of call of duty modern warfare 2</td>
<td>FPS</td>
<td>controlled experiment</td>
</tr>
</tbody>
</table>

¹ Racing games can be categorised as both first and third person avatar games.
² Sports games are difficult to classify in Claypool’s model. Madden NFL Football can be classified as third person avatar or omnipresent.
2004 [3], and the 2006 paper adds no new data, only analysis. Players are set in front of computers running the game "Unreal Tournament 2003" and given a set of tasks. The widely quoted number seems to originate in an experiment where shooting precision was tested. The experiment was run three times by two different players at each latency level. Results from such an experiment are not sufficient to draw any clear conclusions. Other factors were also analysed in this work, but are of little statistical relevance due to the extremely low number of participants. Other numbers cited in [8] are similarly from studies using an extremely low number of participants.

For "omnipresent" games, the most relevant data comes from "The effect of latency on user performance in Real-Time Strategy games". Claypool [7], which in turn uses most of the data from Sheldon et al. [20]. The papers do not establish any clear threshold for latency in such games, simply concluding that 1000ms should be completely safe. Splitting the games in different types of interaction receives significant focus, but differences in skill levels of players are completely ignored. Only two players participate in either study. For the last category of game, third person avatar, the data comes from Fritch et al. [11], which evaluates the performance of two players playing the game Everquest 2 under different conditions.

In "Latency Can Kill: Precision and Deadline in Online Games" [9], Claypool further elaborates on the categories of games and actions. Each action is described by two parameters; deadline and precision. Deadline is the time an action takes to complete, and precision is the accuracy needed by the player. To investigate how sensitivity to latency varies with these two parameters, the authors modified a game, Battle Zone capture the flag (BZFlag), so these parameters could be controlled directly. Each scenario was then played out using computer controlled player avatars called bots. They do not clearly justify that bots are an accurate model for how human players react to latency, neither do they cite any research indicating that this is the case. The hypothesis of a correlation between each of the variables and latency sensitivity was supported, but not strongly.

Team sports games are played in a somewhat different manner to the others mentioned here. Usually you have control of one character at a time, as in a third person avatar game, but you switch character often, depending on who is most involved in the action, as if in an omnipresent game. Nichols and Claypool [16] study the sports game Madden NFL Football, and conclude that latencies as high as 500 ms are not noticeable.

In [18], the authors set up a 12 player match of Unreal Tournament 2003 in a controlled environment. Each player is assigned a specific amount of latency and jitter for the duration of the match. After the match, the players answer a questionnaire about their experience in the game. This study still uses relatively few players, but they are able to conclude that 60ms of latency noticeably reduces both performance and experience of this game. In contrast to [5], [18] finds no effects of jitter. The most probable explanation for this is that jitter is handled well in this game.

Amin et al. [1] also run controlled experiments, but use subjective measures on the FPS game "Call of Duty Modern Warfare 2". Further, they graded participants according to gaming experience. They conclude that the most experienced users are not satisfied with latencies above 100 ms.
Observational Studies

Others have taken different approaches to determining acceptable latency for games. Armitage [2] set up two different servers for the game Quake 3 and monitored the latencies experienced by the players joining the servers. They assume that players will only join servers which have acceptable latency. This methodology might give insight as to how much latency players think is acceptable. However, the study does not take into account which other servers are available, so the results will be heavily influenced by the presence of lower latency servers or lack thereof. Additionally, the latencies players think are acceptable might not be the same as the limit where their performance degrades.

Another interesting approach is that of Chen et al. [5]. They examine an online RPG, ShenZhou Online. By Claypool’s classification this game would be a third person avatar game, and hence be less sensitive to latency than the games discussed earlier. Instead of using a controlled lab environment the authors chose to analyse network traces from an existing, running game. They asked the question: “Does network QoS influence the duration of play sessions?” The Quality of Service (QoS) factors they examined were packet loss, latency and jitter. They hypothesised that if the underlying network conditions affect the players negatively, it should show up in the players’ enjoyment of the game, and hence their motivation to keep playing.

Between 45 and 75 ms RTT, the authors find a linear correlation between increased latency and decreased game session length. For standard deviation of latency, which would represent “jitter”, the correlation is even stronger. At extreme RTT values or extreme jitter the trend is reversed though, and the authors surmise that there is a group of players who are used to bad connections, and another group of players who keep the game on while not really paying attention. These results indicate negative impact of latencies much lower than the 100ms mentioned in earlier literature. Session length as indicator for player satisfaction is an ingenious approach, but the chain of effect from network latency to session length is complicated, and there is significant room for hidden variables.

Dick et al. [10] uses two separate methods to investigate the question. International Telecommunication Union defines what they call an “impairment scale” [13], which defines a rating called ” Using the Mean Opinion Score (MOS) [13] ratings, they ran an online survey asking people how much delay would cause each level of impairment. Players reported they could play "unpaired" at up to about 80 ms, and "tolerable" at 120 ms for most games. Further, the authors ran a controlled experiment testing different latencies. Their results show large differences between games, with the most sensitive game "Need for Speed Underground 2" showing impairment even at the lowest tested delay of 50 ms.

Cloud Gaming

Latency in cloud gaming is much less studied than for networked games. Jarschel et al.2011 [14] test players’ subjective experience of varying network latencies and amount of packet loss in cloud gaming. Using a setup where the gameplay was transferred over a network mimicking the cloud scenario, the authors introduced the varying Quality of Service (QoS) parameters and asked their 48 participants how they liked the service in each scenario. They concluded that a latency of 80ms was noticeable in the fastest-paced game. In all games, packet loss in the stream from server to client was extremely detrimental to the experience. However, their
conclusion can only be used to put an upper bound on latency sensitivity in cloud gaming, because they conducted no experiments between latencies of 0ms and 80ms.

4 Discussion

Based on our review of the field and the findings retrieved from primary sources, we further discuss the implications and general ideas that can be gained from current research.

Limitations in the Classification

The game classification in [8] does not reflect the full spectrum of games, as games in each type can have significantly different latency requirements. In the research summarised in this article, conclusions about noticeable latency in third person avatar games range from 45 ms [5] to over 500 ms [8]. Different games in the same category have very different modes of interaction. Third person avatar games can have very different ways of interaction with the environment. Some are 3D graphical representations of very abstract game mechanics, where relative positions and timing are almost irrelevant, while others are highly detailed simulations of realistic physics where small changes in position and timing make for large changes in the outcome of actions.

For first person avatar games, the differences are often in playing style. Some games focus on tactical movement and cover (for example Metal Gear Solid), while others rely purely on reaction time and precision motor skills (for example Unreal Tournament). It is likely that the last category has stricter latency requirement than the first. No studies found in this survey address these questions.

The distinction between these types of games is also blurred by the fact that many games allow multiple points of view. Role playing games and racing games commonly support both first person and third person view modes for the same game; leaving it up to the user to decide.

Even the status of omnipresent games is not entirely clear. Some are clearly slow-paced and highly strategic, but not all. The well known game Star Craft is one of the most popular games in this category. The papers cited here all seem to agree that this means it should not have strict latency requirements. However, this game is played as a professional sport in parts of the world. In these matches players take around 400 individual actions per minute. This equates to 150 ms per action [15]. A hypothesis that players playing on this level require latencies at least lower than the time between actions would be interesting to investigate.

Types of Input

Users control games in very different ways. FPS and RPG games are sometimes controlled with a handheld controller using analogue joysticks and buttons and in other cases with a keyboard and mouse. Many games have support for both types of input. None of the articles we have surveyed mention which type of input is used, though it is more than likely that this affects the sensitivity of the game. Other games may even have specialised control hardware, such as steering wheels for racing games or digital, or arcade style, joysticks for retro games.
Skill Level
Different players might have different requirements; more skilled players will probably notice latency earlier than amateurs, because they will be more aware of details of the game. On the other hand, experienced players might also have more experience in compensating for and dealing with latency while playing. Players who are used to low latency might react sooner than those who usually play over high-latency links and are used to this situation. None of the studies discussed so far supply strong support for or reject such hypothesis.

Player Assumptions
Most games report latency to players during playing, and all allow for checking this information. Only one source [10] attempts to compare player expectations to their actual experience in games, and finds indications that they are very different. Among the other studies, some partly measure player expectation, such as those that measure session length and connection numbers to a running server. Others attempt measuring the actual effects of latency, but none make it clear how they avoid introducing effects of player expectation.

5 Conclusions and Future Work
This paper has investigated acceptable latency for games. Current research is mostly inconclusive about latency requirements for networked games. In general, it seems that 60 ms [18], or even 45 ms [5] are better estimates at how much latency is acceptable in the most fast-paced games than the traditionally quoted 100ms value. Furthermore, there are no clear, consistent results available and the diversity in game scenarios make comparisons challenging. Studies suffer from uncontrolled environments or very limited numbers of participants. Studies using both a controlled environment and a number of participants that is large enough to do statistical analysis would do a lot to clarify the situation.

Towards a New Classification
Current classification of games is based on genre distinctions used in game design and criticism. These distinctions do not map neatly to what we need to distinguish on a technical level. Considering the limitations of the current classification of games, it is clear that from a technical perspective an updated classification is required. Some factors of this classification present themselves from the work surveyed here.

- Spatial precision describes how much precision in input affects the outcome of an action. This can be the exact angle of the joystick or the distance moved on a mouse.
- Temporal precision describes how much timing of user actions affects the outcome of an action.
- Input type describes how users interact with the game, both in terms of physical devices in use and how these devices control gameplay.

These parameters are not immediately obvious from looking at the game, but they are objective results of the game code. Further research is required to establish testing protocols for these parameters, as well as further elaboration of the model.
Cloud Gaming

Even in the most studied case of networked applications, conclusions are highly diverging. Cloud gaming, the newer case, has barely been studied at all. Current work also lack input from important fields such as neuropsychology in designing the investigations. However, the topic of latency in games is relevant and will influence potentially all games utilising some version of network-based interaction. Therefore we suggest this area is worthy of future pursuit. Forthcoming research would strongly benefit from being based in concrete prototype implementations, in order to generate datasets for more precise conclusions.

References


