Journal of Physics: Conference Series

PAPER • OPEN ACCESS

Research on Latency Problems and Solutions in Cloud Game

To cite this article: Hanlin Sun 2019 *J. Phys.: Conf. Ser.* **1314** 012211

Research on Latency Problems and Solutions in Cloud Game

Hanlin Sun

Shanghai University, Shanghai, 200000, China

Corresponding author's e-mail: angela@cas-harbour.org

Abstract. In recent years, the concept of cloud game remains a hot spot in both academic and game industry, proved by a large number of related research papers and the emergence of various cloud gaming platforms. Cloud game has competitive advantages over other traditional games. However, cloud games are more susceptible to latency than traditional online games because the data of cloud games are processed in remote servers. This paper shows the researches and progress in cloud gaming, analyzes the main aspects that may influence cloud gaming latency and proposes feasible methods to improve these problems.

1. Introduction

After developing for almost twenty years, computer games are becoming increasingly complex. Most modern games are computationally and graphically intensive, such as Assassin's Creed Odyssey, Total War, and Warhammer. To play these games, players need a powerful multi-core CPU and high-end graphics card to run these resource-demanded programs. What's more, it is necessary for players to upgrade their computer hardware continuously since computer games are becoming more and more complex. As a result, users are often restricted by the need for a well-equipped computer and cannot play games at any time and anywhere.

Cloud game offers an alternative solution to these players. Based on cloud computing, games are running on powerful remote servers, all things that players need to do is to receive the data stream from the server and sending commands back to build interactions. In this way, less computation and graphics processing capability is required for users' personal devices and therefore even mobile devices can run computational and graphics demanding games. Thus, players do not need to concern about hardware problems any longer[1]. Given these advantages for game developers and customers, cloud game has been considered as a potential paragon for game development.

The basic idea of the cloud game is to render video games on the cloud and deliver stream-encoded game scene via network transmitting, which means that when players send a signal to the cloud server to control characters' action, they can not see the results until the cloud server makes a response, processes the control order and then renders a new game scene, and eventually sends the screen to the player. The problem with this is that players cannot get a real-time response from the game, and the server needs some time to deal with orders from the clients. A large amount of data will be created during this process. Therefore, the gaming experience of cloud game players is extremely susceptible to response latency[2].

Latency is a common issue for traditional online games. Taking World of Warcraft as an example, game scenes are rendered in players' personal computer, therefore, the impact on the network can be mitigated by using delay compensation techniques[3], such as dead reckoning[4]. However, such techniques cannot be adapted to cloud games since these techniques require game state information, which is not accessible for the player in cloud game structure. In addition, the amount of data in cloud games is much bigger than that of the traditional C/S model online games. Therefore, cloud game is more latency-sensitive than traditional online games.

2. Literature Review

A research team from Taiwan National University built a model to test whether a game is suitable and friendly for transplanting to cloud game platform[2]. By investigating the impact of latency on players' gaming experience, they draw a conclusion that not all games are cloud-gaming friendly, especially for games with dynamic screen changes and high frequency of getting input from users. In their paper, they established a model to predict whether one kind of game is cloud-gaming friendly based on testing latency's impact on players' mood. Their work is a milestone for the development of cloud-game theory but has not explained latency problems and impacts explicitly. In Victor Clincy and Brandon Wilgor's paper, they mainly focused on the network conditions: Packet loss and Round-trip response time[5]. It can be seen that many works have been done to understand the importance of latency's impact on cloud games, but most of them were focused on network systems and few have investigated other potential factors that cause latency problem.

This paper reviews the current directions and research development on studying latency problem of cloud game. The fundamental purpose of this study is to provide new ideas and inspirations to scholars focusing on cloud game in future research.

3. Elements that Contribute to Latency

Influential factors impact on the cloud game latency can be divided into three parts:[6]

- Network delay (ND): Total times required for delivering the command of the player to the server and send back to the computer monitor of the client from the server.
- Playout delay (OD): Different duration of the client receives the encoded form of a frame and the frame is decoded and presented on the screen.
- Processing delay (PD): Different duration of the server receives players' command (from the thin client) and it responds with a corresponding frame after processing the command.

The overall response latency is called Response Delay (RD), and RD=ND+OD+PD.

3.1. Video codec

In the structure of cloud game, delivering powerful server capture, compressing, and rendering the game scene to thin clients do not need strong computational capabilities; and most cloud gaming platforms are using video stream to transmit data.

Currently, H.264 has been widely accepted by most cloud gaming platforms and extensively utilized in video recording and editing. For instance, OnLive is a typical cloud gaming platform that uses H.264 codec. However, in the case of mobile cloud gaming, H.264 is not an efficient coding form anymore. The reason for this is that mobile devices deliver huge amounts of data includes high quality and high-frame-rate graphics under stringent latency requirements. Thus, in most cases, mobile players have to ensure that they have a high-bandwidth network condition. Compressing game image that for transmission is an essential means to mitigate the requirement of bandwidth network. Thus, most of the mobile players use H.264 as the default-coded format.

In most cases, significant differences exist between each graphics frame of graphics-intensive and fast-action games. These differences lead to the inefficiency of applying video codec for compressing game images[7]. For those games with few scene changes, such as SLG game, the correlation of each frame graphic is big; simply using video codec in-game image compression may send repeated information to the server. For solving this problem, a layer-coding approach was proposed to separate the game image into two layers: base layer (contain original image information) and graphics enhancement layer, which contains graphics enhancement instructions: light\map rendering, shading command, reflection computations, etc. This approach uses relatively little computation and rendering resource on clients. The consequent result of this approach is that the client rendering computation and transmission bit-rate can be reduced, and the transmitting data over the network can be decreased significantly as well.

Improvement in video codec aims at diminishing Playout delay (OD) by augmenting the efficiency of decoding thin client, and eventually shortens the time needed for showing the latest game scene rendered by the distant server. Using layer-coding can make both the coding and decoding process faster and more efficient than H.264. This concept offers a fundamental theoretical framework for developing a new encoder that provides an advanced QoE (Quality of Experience) in cloud gaming experience.

3.2. GPU virtualization

Although cloud game provides games with enhanced quality to its users in a completely new way, it is intractable for the company to control their hardware cost while guaranteeing players' gaming experience. For example, OnLive was in financial difficulties because of their plan to improve their players' experience. However, the low-level configuration will certainly bring unsatisfying experience to cloud game players, even worse, they will refuse to play cloud games in the future. Powerful and high-configuration hardware has drawbacks as well. In the traditional mode, one client has at least a server, this structure generates considerable cost[8]. On the other hand, given the fact that each game has its own performance requirement standard, in the case of the 2D game, it is a waste of resource and money to prepare a super-powerful server, which can run high-quality graphic ACT/FPS game smoothly.

Virtualization plays a crucial role in integrating game resources and decreasing hardware cost. By using virtual machines such as VMWare, it enables multiple cloud games to share servers' hardware resources while keeping isolated. So far, CPU, network interfaces, and storage virtualization technique have been developed into mature conditions. However, GPU is an exceptional case. As pointed out by previous scholars, GPU virtualization has a high probability of causing higher latency and more extended response time. A newly developed method of GPU virtualization may help to make some improvement to this issue. In the past, the virtual machine's graphics rendering capability cannot run GPU-intensive games smoothly. The efficiency of CPU's graphics ability is low, while GPU is good at handling the computation of massive matrix data, which can strongly improve the graphics processing speed. The problem with this is that there is no GPU algorithm matured enough to cope with virtual machines. Thus, the development of graphics capability for the virtual machine has been limited.

Previous researchers proposed a new GPU algorithm named "vGPU" for the virtual machines[10]. The basic idea of "vGPU" is to divide GPU computation resources into many pieces and invite CPU in graphics processing to compensate for the loss of GPU computing resources. By sharing GPU computation resources among virtual machines, their graphics capability will be enhanced to some extent. Another group of scholars managed to build an original cloud game system by using vGPU techniques and a cloud gaming platform called GamingAnywhere[11]. The result shows that it indeed lower the cloud game latency by improving processing ability in each virtual machine, but its effect is insignificant. "vGPU" can only fit the basic requirement of cloud game, but its impact on the progressing of cloud gaming platform is marginal.

Virtualizing GPU provides multiple benefits to virtual machine. To be specific, virtualizing GPU increases the server's game concurrency ability and enhances virtual machine's computation capability. Moreover, GPU virtualization decreases virtual machine's processing delay and reduces hardware cost. Experiment also shows that installing GPU in virtual machine can improve coding efficiency. Problem with GPU virtualization is one of the biggest obstacles that have impeded cloud game from achieving satisfying latency and response time. Therefore, a practical algorithm that can improve virtual machine's graphics capability prominently is an urgent need. The author believes that it is necessary and essential for future research on cloud gaming platform to put emphasis on GPU virtualization.

3.3. Network latency

Network latency is an inevitable problem in both traditional online game and cloud game. Especially for cloud game, based on cloud techniques, cloud game servers are distributed across geographical locations, so choosing the server with the shortest latency for each player became a difficult task. In traditional online games, to solve network latency problem, experts invited distributed game server architectures and extensible network structure[12]. However, in the case of cloud game, this method cannot be applied directly. Each cloud game server provides a large number of games to its users and each game has different requirements for network resources. Choosing the server with the shortest network latency does not mean it is the fastest one because each server's computation capability may be different due to virtual machines. Game is running on cloud servers instead of personal computers. Thus, it would be difficult to make any compensation on latency caused by network delay, for example, dead reckoning is a common method used to address latency. The huge amount of data transmitting should also be considered in network.

These problems generate two topics as the following: how to choose a proper server in cloud games, and how to improve network conditions. The emergence of the 5G protocol may be a practical way to improve network conditions. Research shows that 5G protocol can bring low latencies for those graphics and data-intensive applications, such as virtual reality and augmented reality[13]. This new technology is useful for cloud games too. In the architecture of cloud game, almost all computations are done in the center of the cloud – the powerful server. However, since the powerful server is limited by network speed and the scale of data amount, not all data can be sent to the client on time. In some frame-intensive game with high graphics quality, congestion may happen and bring the players negative emotions. Therefore, the high transmitting speed of 5G network plays an important role in improving this situation[14]. Even though the coverage of 5G network has not extended to a broader era yet, it still can be used on the architecture of cloud game. As a result, the network latency of the cloud gaming platform developed based on the 5G protocol will reduce significantly.

3.4. Models of computation

A majority of cloud gaming platforms are facing the problem of high round trip time due to the cloud data center is distant from the clients. This problem directly leads to comparatively higher network latency. When a large scale of users are requesting services, the data center's defect of needing huge bandwidth consumption and lacking game concurrence ability was conspicuous. Immature GPU virtualizing techniques also impede cloud games' development.

Lin Li's study shows that Edge computation can be invited in cloud game architecture to lower the latency[15]. Edge computation is a new computation model developed with the emergence of IoT (Internet of Things) and 5G protocol. In this model, computation is done at the edge of the network instead of the center (in C/S model, it's a server). Edge computation model manipulates the edge node's network proximity to solve the latency problem in cloud games, like a distributed computation system, which overcomes the high bandwidth consumption in the cloud center.

The update of the computation model can make up for the overall response delay and diminish network latency prominently. Lin Li and his team have cultivated a new cloud game frame, which is called Gaming@Edge. They investigated a new mechanism using graphic stream based Game-on-demand, games are running on the server, but the server will intercept client's demand of rendering and sending the graphic stream to the client, the client will be responsible for finishing graphic command. This creative idea managed to divide game scene's rendering process from game logic's computation, manipulate Edge node's computation ability and make full use of client's render ability. Eventually, this mechanism greatly reduced system latency and improved server's game concurrency. Combining with 5G protocol, Edge computation in cloud games has a prospecting future in both the academic research and industrial area.

4. Conclusion

Currently, cloud game has become increasingly popular. It is foreseeable that cloud game will continue to develop under the trend of chasing the cloud computing technique. However, unsolved problems still exist in cloud game such as hardware limitation, network speed, and GPU virtualizing. These issues are the main reasons that hindered cloud game from becoming more popular and widely accepted by players. Among these problems, the latency problem is the most important one that has substantial impact on the development of cloud gaming platform.

This paper describes the latency problem of cloud gaming platform and primary factors contribute to the overall latency (or response delay) and reviewed current progress on research focused on cloud game. The author proposed a possible solution for each factor and pointed out three prospecting research directions, which is solving cloud game latency problem, 5G protocol application, and cloud game Edge Computation model. Breakthroughs in these fields of research will improve cloud game’s latency problem and breathe new life into cloud game’s development, especially for the 5G network application. This new technology has been put into use in small scale by many countries. Since the maturity level of the 5G network’s techniques is continuously increasing, it is foreseeable that 5G network will replace 4G network in the following years. In the case of cloud game, integrating 5G network and Edge computation with cloud game frame will make a significant improvement in the problem of network delay.

References

[1] Ross, Philip. E. Cloud computing’s killer app: gaming. Spectrum IEEE, 2009, 46(3): 14.

[2] Lee, Yeng-Ting, Chen, Kuan-Ta, et al. Are all games equally cloud-gaming-friendly? An electromyographic approach. In Proceedings of the 11th Annual Workshop on Network and Systems Support for Games (NetGames’12). IEEE, 2012.

[3] Yahn W. Bernier. Latency compensating methods in client/server in-game protocol design and optimization. Published in the 15th Game Developers Conference, Mar 2001.

[4] Pantel, L., Wolf, L. C. On the suitability of dead reckoning schemes for games. In Proceedings of the 1st Workshop on Network and System Support for Games, NETGAMES 2002, Braunschweig, Germany, April 16-17, 2002.

[5] Clincy, V., Wilgor, B. Qualitative evaluation of latency and packet loss in a cloud-based games. GSTF Journal on Computing (JoC), 2013,3(1):473-476.

[6] Shea, R., Liu, J. On GPU pass-through performance for cloud gaming: experiments and analysis. In Proceedings of IEEE 12th Annual Workshop on Network and Systems Support for Games(Net Games) 2013:1-6.

[7] Chen Kuan-Ta, Huang Chun-Ying, Hsu Cheng-Hsin. Cloud gaming onward: research opportunities and outlook. Published in 2014 IEEE International Conference on Multimedia and Expo Workshops (ICMEW)

[8] Dai Jia-wei, Bai Guang-wei, et al. Performance analysis of cloud gaming in GPU virtualization environment. Journal of Chinese Computer Systems, Feb 2018

[9] Dowty M, Sugerman J. GPU virtualization on VMware’s hosted I/O architecture. Published in ACM SIGOP Operating Systems Review, 2009

[10] Li J., Li C., et al. vGPU: a real time GPU emulator: U.S Patent, us8711159[P] 2014-4-29.

[11] Chun-Ying Huang, Cheng-Hsin Hsu, et al. Gaming anywhere: an open cloud gaming system. In Proceedings of the 4th ACM Multimedia Systems Conference (MMSys’13). ACM, New York, NY, USA, 2013:36-47.

[12] Lee, Kang Won, Ko, Bong Jun, Calo, Seraphin. Adaptive server selection for large scale interactive online games. Computer Networks, 49(1), 2005:84–102

[13] Schmoll, Robert Steve, Pandi, Sreekrishna, et al. Demonstration of VR/AR offloading to mobile edge cloud for low latency 5G gaming application, Published in 2018 15th IEEE Annual Consumer Communications & Networking Conference (CCNC), 2018:1-3

[14] Qian Clara Li, Niu Huaning, et al. 5G network capacity: key elements and technologies, IEEE Vehicular Technology Magazine 9(1), 2014: 71-78

[15] Lin Li, Xiong Jinbo, et al. Gaming@Edge: a low latency cloud gaming system based on edge nodes. Journal of Computer Applications, 2019