CLOUD GAMING ONWARD: RESEARCH OPPORTUNITIES AND OUTLOOK

Kuan-Ta Chen[1]*, Chun-Ying Huang*[2]*, and Cheng-Hsin Hsu*[3]

[1]Institute of Information Science, Academia Sinica
[2]Department of Computer Science and Engineering, National Taiwan Ocean University
[3]Department of Computer Science, National Tsing Hua University

ABSTRACT

Cloud gaming has become increasingly more popular in the academia and the industry, evident by the large numbers of related research papers and startup companies. Some public cloud gaming services have attracted hundreds of thousands subscribers, demonstrating the initial success of cloud gaming services. Pushing the cloud gaming services forward, however, faces various challenges, which open up many research opportunities. In this paper, we share our views on the future cloud gaming research, and point out several research problems spanning over a wide spectrum of different directions: including distributed systems, video codecs, virtualization, human-computer interaction, quality of experience, resource allocation, and dynamic adaptation. Solving these research problems will allow service providers to offer high-quality cloud gaming services yet remain profitable, which in turn results in even more successful cloud gaming eco-environment. In addition, we believe there will be many more novel ideas to capitalize the abundant and elastic cloud resources for better gaming experience, and we will see these ideas and associated challenges in the years to come.

Index Terms— Computer games, cloud computing, interactive applications, performance optimization, networked systems

1. INTRODUCTION

The increasing prevalence of cloud infrastructures provide abundant computing, storage, and communication resources in a cost-effective, reliable, elastic, high-performance, low-maintenance manner. These cloud computing resources may be leveraged by various applications, and among them the resource-hungry *computer games* have been recognized as the killer application for cloud computing [1]. In cloud computing, powerful cloud servers render, capture, compress, and transmit game screens to thin clients running on relatively less-capable computing devices. The thin clients decode and display game screens to players. Gamers' inputs are also collected and sent by thin clients to cloud servers in real time. Cloud gaming was predicted to be the fastest growing game industry sector [2], and has in fact accumulated tremendous momentum in both the industry [3–5] and the academia [6, 7]. For example, CloudUnion [5] reports to have 20 million subscribers in China, which we believe has gone beyond the *critical mass*, demonstrating the bright future of cloud gaming.

Delivering good cloud gaming experience, however, is no easy task due to the (geographically) distributed nature of cloud infrastructures, the best-effort Internet, the strong interactiveness of computer games, and high expectations of players. In particular, players concurrently demand for both fast responsiveness and high-definition game screens, which are already challenging to cloud gaming providers. Moreover, providers must deliver such cloud gaming experience in a cost-effective, scalable, and error-resilient way, which further complicates the task. These new challenges open up a full spectrum of research opportunities, which are of great interests to research communities. In this position paper, we share our views on the key cloud gaming research opportunities. We firmly believe that addressing these research problems will turn cloud gaming to even bigger success.

2. RESEARCH OPPORTUNITIES

2.1. Game Integration

One main part of cloud gaming technology is to take the rendered screens of computer games as the input, encode the screens, and transmit the coded screens to the client for display. How the rendered game screens are taken (or captured) for video encoding remains an ad hoc fashion. When the game source code is available, we can directly modify the game's rendering engine and instruct the engine to send out each game screen update once the rendering is finished. Otherwise, a cloud gaming platform may need to rely on certain system hacks to 1) intercept the event whenever a screen rendering is finished, and then 2) copy the rendered screen to the input buffer of a video encoder, which may or may not involve inter-process communications and context switches. The main disadvantage of the approach is that it is highly system- and game-dependent, as any changes in the game rendering engine and its underneath libraries (such as Microsoft DirectX) may invalidate such system hacks. Another flaw is that it may incur performance penalties as the interceptions

and memory copies tend to introduce processing overhead and system instability.

In view of the above issues, we consider that the integration of games and cloud gaming platforms is worthy of serious investigation. We believe that a framework that enables the integration of games into cloud gaming platforms in a platform-independent way will be highly desired to enable a scalable, rich cloud gaming eco-environment, as with such a framework, a game would be easily "plugged" into any cloud gaming platform without extra efforts. The integration frameworks need to take account of 1) the common practices of rendering engine design, 2) how the engines are utilized by game developers, and 3) how the platform handles the captured game screens. Thus, it is expected that developing such frameworks would be joint efforts between the game developers and cloud gaming system designers.

2.2. Video Codec

Currently H.264 is the de facto video codec for cloud gaming, e.g., it has been adopted by OnLive, the first commercial cloud gaming operator. Although H.264 is indeed a decent general-purpose video codec, it is shown that H.264 may not be the most efficient codec for cloud gaming [8, 9]. One of the reasons is that different games may feature different graphic styles with drastic variety: some feature realistic game scenes, and some feature cartoon-like style with a succinct manner. Meanwhile, some codecs can better handle video with particular graphical traits than others. Another reason is that the current design of H.264 does allow us to do *cross-layer optimization* across graphics rendering and video coding. For example, a layered-coding approach was proposed [8] to separate a game frame into a baseline layer and an enhancement layer, where the enhancement layer contains some graphics-enhancing instructions (such as lighting and shading commands) to the rendering engine on the client. This approach utilizes relatively little computation and rendering resources on the client to exchange for a reduced workload on the server (on graphics rendering and coding) and a reduced amount of transmitted data over the network. This demonstrates how the cross-layer optimization can be achieved if the rendering engine (as part of games) and video coding (as part of cloud gaming platforms) can seamlessly work together to yield an overall better efficiency and provide a better Quality of Experience (QoE).

2.3. Virtualization

From the experience of OnLive, we have learned that business operations largely decide the success and failure of a cloud gaming vendor. One known issue of OnLive is that its server is not well scalable in terms of the number of game instances running on a physical server [10]. One would expect that with the support of modern para-virtualization technologies a server can run a large number of game instances on a single server by using virtual machines such as Xen, VMWare, and KVM. Our experience reports that a number of issues limit the scalability of gaming instances on servers. Notably, the GPU virtualization technology still has huge space for improvements before multiple GPU-intensive games can smoothly run in their respective virtual machines. As a common practice, game developers assume that their games completely own the GPU and GPU memory and plan the GPU memory in an exclusive way to reduce graphics loading time (from the disk) and to speedup rendering frame rate. Therefore, GPU virtualization is essential so that each of the game instances can be provided its own virtual GPU and GPU memory as though it owns the GPU and GPU memory exclusively. At the same time, GPU rendering is memory-intensive such that GPU memory virtualization would largely decrease the rendering and display performance, which together make the GPU virtualization, especially for cloud gaming support, a challenging research direction.

2.4. User Interface

When it comes at mobile cloud gaming [11], the design of user interface plays a critical role in affecting user experience especially if the streamed games were not originally designed for mobile use. More specifically, most PC games rely on the combinations of keyboard and mouse as input devices for gaming controls, whereas mobile devices only provide touch interfaces. So far, there is no straightforward mapping from keyboard and mouse inputs to touch events, and therefore how to provide a natural user interface on mobile devices for non-mobile games constitutes another research challenge. The state-of-the-practice solution is to manually design mobile interfaces in such scenarios, but this is certainly not scalable considering the number of non-mobile games that are potentially to be played by mobile users. Thus, mechanisms for (semi-)automatic mapping between the non-mobile and mobile user interfaces are highly demanded for mobile cloud gaming.

2.5. QoE Measurement and Modeling

Unlike multimedia content such as images and videos, game play is a dynamic and interactive process, where users' experience can vary over time and the game contents continuously change depending on what game inputs have been received. Therefore, measuring the QoE provided by a cloud gaming system is a challenging research topic. In the simplest setting, we can ask players to report their gaming experience for a whole game session; however, this measurement is subject to the primacy and the recency effect [12] as people's short-term memory is quite limited and cannot remember every detailed perception during the game play. Common QoE measurement methods such as SSCQE and DSCQE [13]

are not directly applicable to gaming scenarios because players have to focus on game play and normally both of their hands are occupied (holding the mouse, hitting keys, sliding on touch screens, and/or holding the mobile device). Paired comparison [14] has been proposed to measure the QoE during gaming but it also captures segment QoE rather than continuous QoE and the number of trails grow quadratically with the number of stimuli, which would make the user study infeasible when the number of stimuli is larger than 10. Physiological methods were proposed as a solution to continuously monitor players' perceptions during game play [15], but the mapping between players' bio-signals, such as electromyography measurements and heart rates, and their emotions is non-trivial and remains to be investigated.

Despite of the difficulties in capturing players' continuous gaming experience, the QoE models, which describe the relationship between system/network parameters and players' gaming experience, w ould be foundations to subsequent developments of QoE-aware cloud gaming systems. For example, such a system would be able to automatically adjust the system parameters, such as the video coding bitrate and frame sampling rate, according to the configuration and environment parameters, such as network bandwidth, network delay, and display size used by the player, in order to provide a more satisfactory experience. The adaption of systems can include one-time, static decisions such as the selection of cloud game servers, and continuous adjustments of system parameters, such as the video coding bitrate. The most common strategies for QoE management are discussed in the remaining sections.

2.6. Server Selection

If the cloud gaming servers are distributed across geographical locations, whenever a user attempts to log into the system and starts playing games, a server selection problem would naturally arise [16]. Although this problem has been well studied in the field of online games [17], the selection of cloud gaming servers remains to be explored because its distinct features (compared to online gaming): 1) A large number of games are normally provided by a cloud gaming platform, where each game may have very different resource requirements; 2) the computation resources of each server are normally heterogeneous due to legacy issues and virtualization; and 3) network delay in cloud gaming is more critical than that in online gaming because cloud gaming clients do not possess game state information, and thus there are much less opportunities for performing delay compensation, such as dead reckoning [18].

Intuitively, a server with the shortest network delay to a player should be chosen to be provided to the user. However, a number of potential issues would make the problem much more complicated, such as: 1) the server may be already overloaded by serving other players; 2) the server may have relatively less resources such that a graphics-intensive game cannot run smoothly on it; 3) if the operator supports many games, say, hundreds of games, probably not all of the games are available on each server in order to save disk space; 4) the player may choose a multi-user game so that we need to consider also the network locations of the other players who participate in the same game; and 5) the player may choose an online game, so it is better to provide him a server which provides the shortest overall network delay (i.e., the network delay between the client and the cloud gaming server and that between the cloud gaming server and the online game server). Despite the potential complexity and high dimension of server selection, if addressed well, it will largely affect how smooth the subsequent game play is as the network delay is one of the dominant factors in QoE models.

2.7. Parameter Adaptation

There are a bunch of parameters that can be configured in the run time on a cloud gaming server in order to keep a balance between workload and gaming experience given the constraints of the environment, such as the network conditions and the player's device capabilities. For example, we can easily tune down video quality when packet loss is significant; however, this may significantly lower players' gaming experience. Instead, we may sample the game screens in a lower rate in exchange for a higher video quality while keeping the overall bandwidth usage intact and make the players happier. This strategy may succeed or fail depending on a large number of factors including the display size of the client and the genre of the game being played [15]. To achieve a smart parameter adaptation that maintains the tradeoffs among various parameters, we would require a sophisticated QoE model that keeps track of players' perceptions with parameter configurations and devises learning or control-theoretic algorithms that control the parameter settings while taking account of the dynamically changing environment, e.g., network delay and available bandwidth may change anytime.

2.8. Resource Scheduling

Different games have different workloads in terms of CPU, GPU, and memory usage. Furthermore, one single game can also have different workloads depending on the stage or scene currently played. Thus, normally it is better to mix games with different resource requirements rather than running multiple instances of a game on a physical server to make a more efficient allocation of server resources. For example, a server may not be able to run three graphics-intensive games at the same time (due to limits in GPU capability), but it may be able to simultaneously run one graphics-intensive game and five non-graphics intensive games on the same server. This multiplexing strategy, if applied properly, can largely increase the overall service coverage while maintaining the desired QoE provisioning levels. However, this line of CPU-GPU-memory

co-scheduling research issue [19] remains an unsolved research challenge in the community.

3. CONCLUSION AND OUTLOOK

Cloud gaming is getting increasingly popular, e.g., CloudUnion [5] has too many subscribers compared with its current infrastructure, and an admission control algorithm was proposed [20] to alleviate the long waiting time. To turn cloud gaming into an even bigger success, there are still many challenges ahead of us. In this paper, we share our views on the most promising research opportunities for providing high-quality and commercially-viable cloud gaming services. These opportunities span over fairly diverse research directions: from very system-oriented game integration to quite human-centric QoE modeling; from cloud related GPU virtualization to content-dependent video codecs. We believe these research opportunities are of great interests to both the research community and the industry for future, better cloud gaming platforms.

The current success of cloud gaming is only the tip of the iceberg, and many creative and new ideas of leveraging the abundant and elastic cloud resources for better interactive user experience will surface soon. For example, mobile devices may display high-quality game screens rendered in one or multiple distributed cloud servers [8, 9], which were not possible on resource-constrained mobile devices. While these novel ideas will unleash the potentials of cloud computing, we also expect to face new and exciting research challenges in the years to come.

References

[1] P. Ross, "Cloud computing's killer app: Gaming," *IEEE Spectrum*, vol. 46, no. 3, p. 14, March 2009.

[2] "Distribution and monetization strategies to increase revenues from cloud gaming," http://www.cgconfusa.com/report/documents/Content-5minCloudGamingReportHighlights.pdf.

[3] "Onlive web page," http://www.onlive.com/.

[4] "Gaikai web page," http://www.gaikai.com/.

[5] "Cloudunion web page," http://www.cloudunion.cn.

[6] R. Shea, J. Liu, E. Ngai, and Y. Cui, "Cloud gaming: Architecture and performance," *IEEE Network*, vol. 27, no. 4, pp. 16–21, July-August 2013.

[7] D. Mishra, M. E. Zarki, A. Erbad, C. Hsu, and N. Venkatasubramanian, "Clouds + games: A multi-faceted approach," *IEEE Internet Computing*, February 2014, accepted to appear.

[8] S.-P. Chuah and N.-M. Cheung, "Layered coding for mobile cloud gaming," in *Proceedings of ACM International Workshop on Massively Multiuser Virtual Environments (MMVE'14)*, March 2014.

[9] S. Shi, C.-H. Hsu, K. Nahrstedt, and R. H. Campbell, "Using graphics rendering contexts to enhance the real-time video coding for mobile cloud gaming," in *Proceedings of ACM International Conference on Multimedia (MM'11)*, November 2011, pp. 103–112.

[10] "OnLive lost: How the paradise of streaming games was undone by one man's ego," http://www.theverge.com/2012/8/28/3274739/onlive-report.

[11] C.-Y. Huang, C.-H. Hsu, D.-Y. Chen, and K.-T. Chen, "Quantifying user satisfaction in mobile cloud games," in *Proceedings of ACM Workshop on Mobile Video Delivery (MoViD'14)*, March 2014, pp. 4:1–4:6.

[12] C. W. Mayo and W. H. Crockett, "Cognitive complexity and primacy-recency effects in impression formation." *The Journal of Abnormal and Social Psychology*, vol. 68, no. 3, p. 335, 1964.

[13] N. Lodge and D. Wood, "New tools for evaluating the quality of digital television-results of the MOSAIC project," in *Proceedings of International Broadcasting Convention (IBC'96)*, September 1996, pp. 323–330.

[14] Y.-C. Chang, K.-T. Chen, C.-C. Wu, C.-J. Ho, and C.-L. Lei, "Online game QoE evaluation using paired comparisons," in *Proceedings of IEEE International Workshop Technical Committee on Communications Quality and Reliability (CQR'10)*, June 2010.

[15] Y.-T. Lee, K.-T. Chen, H.-I. Su, and C.-L. Lei, "Are all games equally cloud-gaming-friendly? An electromyographic approach," in *Proceedings of the IEEE/ACM Annual Workshop on Network and Systems Support for Games (NetGames'12)*, October 2012, pp. 3:1–3:6.

[16] S. Choy, B. Wong, G. Simon, and C. Rosenberg, "The brewing storm in cloud gaming: A measurement study on cloud to end-user latency," in *Proceedings of the IEEE/ACM NetGames 2012*, October 2012.

[17] K.-W. Lee, B.-J. Ko, and S. Calo, "Adaptive server selection for large scale interactive online games," *Computer Networks*, vol. 49, no. 1, pp. 84–102, September 2005.

[18] K.-T. Chen, Y.-C. Chang, H.-J. Hsu, D.-Y. Chen, C.-Y. Huang, and C.-H. Hsu, "On the quality of service of cloud gaming systems," *IEEE Transactions on Multimedia*, vol. 16, no. 2, pp. 480–495, February 2014.

[19] V. Gupta, K. Schwan, N. Tolia, V. Talwar, and P. Ranganathan, "Pegasus: Coordinated scheduling for virtualized accelerator-based systems," in *Proceedings of USENIX Annual Technical Conference (ATC'11)*, June 2011, p. 31.

[20] D. Wu, Z. Xue, and J. He, "iCloudAccess: Cost-effective streaming of video games from the cloud with low latency," *IEEE Transactions on Circuits and Systems for Video Technology*, January 2014, accepted to appear.