Received May 30, 2016, accepted June 13, 2016, date of publication August 8, 2016, date of current version November 18, 2016.

Digital Object Identifier 10.1109/ACCESS.2016.2590500

A Survey on Cloud Gaming: Future of Computer Games

WEI CAI[1], (Member, IEEE), RYAN SHEA[2], (Member, IEEE), CHUN-YING HUANG[3], (Member, IEEE), KUAN-TA CHEN[4], (Senior Member, IEEE), JIANGCHUAN LIU[2], (Senior Member, IEEE), VICTOR C. M. LEUNG[1], (Fellow, IEEE), AND CHENG-HSIN HSU[5], (Senior Member, IEEE)

[1]Department of Electrical and Computer Engineering, The University of British Columbia, Vancouver, BC V6T 1Z4, Canada
[2]School of Computing Science, Simon Fraser University, Burnaby, BC V5A 1S6, Canada
[3]Department of Computer Science, National Chiao Tung University, Hsinchu 300, Taiwan
[4]Institute of Information Science, Academia Sinica, Taipei 115, Taiwan
[5]Department of Computer Science, National Tsing Hua University, Hsinchu 300, Taiwan

Corresponding author: C.-H. Hsu (chsu@cs.nthu.edu.tw)

This work was supported in part by the University of British Columbia through the Four Year Doctoral Fellowship, in part by the Natural Sciences and Engineering Research Council of Canada under Grant STPGP 447524, in part by the National Natural Science Foundation of China under Grant 61271182, in part by the Ministry of Science and Technology of Taiwan under Grant 102-2221-E-007-062-MY3, Grant 103-2221-E-009-230-MY2, Grant 104-2221-E-009-200-MY3, and Grant 103-2221-E-001-023-MY2, in part by an NSERC E.W.R. Steacie Memorial Fellowship, in part by an NSERC Strategic Project Grant, and in part by a Discovery Grant.

ABSTRACT Cloud gaming is a new way to deliver high-quality gaming experience to gamers anywhere and anytime. In cloud gaming, sophisticated game software runs on powerful servers in data centers, rendered game scenes are streamed to gamers over the Internet in real time, and the gamers use light-weight software executed on heterogeneous devices to interact with the games. Due to the proliferation of high-speed networks and cloud computing, cloud gaming has attracted tremendous attentions in both the academia and industry since late 2000's. In this paper, we survey the latest cloud gaming research from different aspects, spanning over cloud gaming platforms, optimization techniques, and commercial cloud gaming services. The readers will gain the overview of cloud gaming research and get familiar with the recent developments in this area.

INDEX TERMS Clouds, distributed computing, video coding, quality of service, computer graphics.

I. INTRODUCTION

Cloud gaming refers to a new way to deliver computer games to users, where computationally complex games are executed on powerful cloud servers, the rendered game scenes are streamed over the Internet to gamers with thin clients on heterogeneous devices, and the control events from input devices are sent back to cloud servers for interactions. Figure 1 presents how cloud gaming services work. In the cloud, a cloud gaming platform is hosted on cloud servers in one or multiple data centers. The cloud gaming platform runs computer game programs, which can be roughly divided into two major components: (i) game logic that is responsible to convert gamer commands into in-game interactions, and (ii) scene renderer that generates game scenes in real-time. The gamer commands come from the command interpreter, and the game scenes are captured by video capturer into videos, which are then compressed by video encoder. The command interpreter, video capturer, and video encoder are all implemented as parts of the cloud gaming platform. As shown in this figure, the cloud gaming platform sends the video frames to, and receives user inputs from thin clients used by gamers for playing games. It is a thin client, because only two low-complexity components are required: (i) command receiver, which connects to the game controllers, such as gampads, joysticks, keyboards, and mouses, and (ii) video decoder, which can be realized using massively produced (inexpensive) decoder chips. The communications between the cloud game platform and thin clients are over the best-effort Internet, which in turn makes supporting real-time computer games quite challenging.

In late 2000's, we started to see cloud gaming services offered by startups, such as OnLive [67], Gaikai [27], G-cluster [26], and Ubitus [93]. We also witnessed that Gaikai was acquired by SONY, which is a major game console developer [86]. This was followed by the competition between Sony's PlayStation Now (PS Now) [68] and Nvidia's Grid

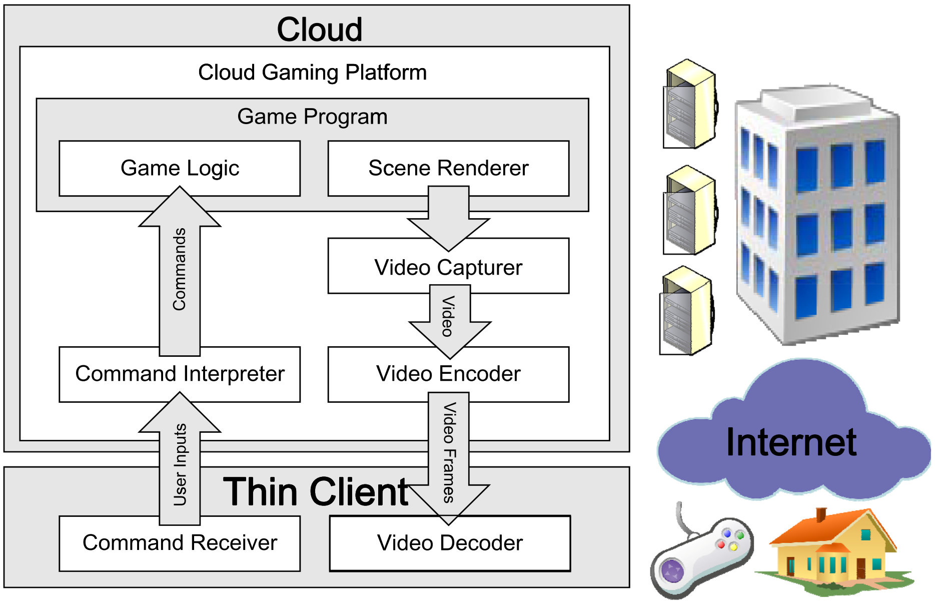

FIGURE 1. Typical cloud gaming services.

Game Streaming Service [65], which further heats up the cloud gaming market. In fact, a 2014 report from Strategy Analytics [75] indicates that the number of cloud gaming users increases from 30 millions in 2014 to 150 millions in 2015. The same report also predicts that other leading game console manufactures will soon join the cloud gaming market.

The tremendous popularity of cloud gaming may be attributed to several potential advantages to gamers, game developers, and service providers. For gamers, cloud gaming enables them to: (i) have access to their games anywhere and anytime, (ii) purchase or rent games on-demand, (iii) avoid regularly upgrading their hardware, and (iv) enjoy unique features such as migrating across client computers during game sessions, observing ongoing tournaments, and sharing game replays with friends. For game developers, cloud gaming allows them to: (i) concentrate on a single platform, which in turn reduces the porting and testing costs, (ii) bypass retailers for higher profit margins, (iii) reach out to more gamers, and (iv) avoid piracy as the game software is never downloaded to client computers. For service providers, cloud gaming: (i) leads to new business models, (ii) creates more demands on already-deployed cloud resources, and (iii) demonstrates the potential of other/new remote execution applications, since cloud gaming imposes the strictest constraints on various computing and networking resources.

Despite the great opportunities of cloud gaming, several crucial challenges must be addressed by the research community before it reaches its full potentials to attract more gamers, game developers, and service providers. We summarize the most important aspect as follows. First, cloud gaming platforms and testbeds must be built up for comprehensive performance evaluations. The evaluations include measurements on Quality of Service (QoS) metrics, such as energy consumption and network metrics, and Quality of Experience (QoE) metrics, such as gamer perceived experience. Building platforms and testbeds, designing the test scenarios, and carrying out the evaluations, require significant efforts, while analyzing the complex interplay between QoS and QoE metrics is even more difficult.

Second, the resulting platforms and evaluation procedures allow the research community to optimize various components, such as cloud servers and communication channels. More specifically, optimization techniques for: (i) better resource allocation and distributed architecture are possible at cloud servers, and (ii) optimal content coding and adaptive transmissions are possible in communication channels.

Third, computer games are of various game genres [19]. These genres can be categorized on the basis of two elements: *viewpoint* and *theme*. *Viewpoint* is how a gamer observe the game scene. It determines the variability of rendered video on the screen. Most commonly seen viewpoints include first-person, second-person, third-person, and omnipresent. First-person games adopt graphical perspectives rendered from the viewpoint of the in-game characters, such as in Counter-Strike. Second-person games are rendered from the back of the in-game characters, so that gamers can see the characters on the screen, like in Grand Theft Auto. Third-person games fix the gamers' views on 3D scenes, projected onto 2D spaces. Modern third-person games usually adopts the sky view, also known as God view. Classic third-person games include Diablo, Command & Conquer, FreeStyle, and etc. Last, omnipresent enables gamers to fully control views on the region of interest (RoI) from different angels and distances. Many recent war games, e.g., Age of Empires 3, Stronghold 2, and Warcraft III, fall into this category. Game *theme* determines how gamers interact with game content. Common themes include shooting, fighting, sports, turn-based role-playing (RPG), action role-playing (ARPG), turn-based strategy, real-time strategy (RTS), and management simulation. Although the *viewpoint* may be restricted by game *theme*, but generally a game genre can be describe by a pair of *viewpoint* and *theme*, such as first-person shooting, third-person ARPG, omnipresent RTS, and etc. Among them, fast-paced first-person shooting games impose the highest scene complexity, which are the most challenging games for cloud gaming service providers. In contrast, third-person turn-based RPG games are least sensitive to delays and thus more suitable for cloud gaming.

Cloud gaming is an exciting research area and the existing literature aims to address several aforementioned challenges. Nonetheless, to the best of our knowledge, there is no comprehensive survey on cloud gaming research. The lack of a central survey of existing literature may delay or even prevent researchers, who are interested in cloud gaming or other remote execution applications, from joining the community. A thorough understanding and exploration of existing academic and industrial research and development can help lead to the building of future cloud gaming platforms. One such advance might come from future games being designed specifically with cloud gaming functionalities and supports in mind. How we accomplish this is still an open question, for example game developers could create cloud gaming aware contexts or even whole new programing paradigms. With this in mind, we carefully connect existing research on solving current challenges together, and come up with a classification system described below.

A. SCOPE AND CLASSIFICATIONS

In the current article, we survey the cloud gaming literature. We first collect representative cloud gaming papers, and group them into several classifications. We emphasize that only a selective set of papers are surveyed, in order to give the readers better understanding on the landscape of the cloud gaming research. Upon selecting the representative papers, we propose a classification system as summarized in Figure 2. More details on the classification system follow.

1) **Cloud Gaming Overview (Section II):** We survey the overview, introductory, and positioning papers on either general cloud gaming, or specialized topics, such as mobile cloud gaming and Game-as-a-Service (GaaS).
2) **Cloud Gaming Platforms (Section III):** We consider papers that construct basic cloud gaming platforms, which support different performance evaluation methodologies. These studies can be further categorized into three groups: system integration, QoS evaluations, and QoE evaluations.
    a) **System Integration (Section III-A):** The fundamental step of cloud gaming research, like many other systems areas, is to put up basic platforms, based on existing tools. We summarize such system integration efforts, which serve as cornerstones of related research.
    b) **Quality of Service Evaluations (Section III-B):** We survey the studies on objective metric evaluations, which algorithmically quantify the system performance, i.e., without subject assessments. Existing papers focus on two types of objective metrics, **Energy Consumption** and **Network Metrics**. The energy consumption is critical to mobile cloud gaming clients, in order to prolong the precious battery life. There are several network metrics affecting the gamer experience, and interaction latency is a representative network metric. The interaction latency refers to the time difference between a gamer input and the corresponding game scene update on the client computer. Because gamers are highly sensitive to interaction latency [19], its measurement methodologies draw a lot of attentions in the literature.
    c) **Quality of Experience Evaluations (Section III-C):** We discuss the papers on subjective metric evaluations, which are based on user studies, where subject gamers give opinion scores to their cloud gaming experience. Conducting user studies is inherently expensive and tedious, and thus most QoE studies attempt to analyze the relationship between the QoS and QoE metrics. The resulting models may in turn be used to optimize cloud gaming platforms.
3) **Optimizing Cloud Gaming Platforms (Section IV):** We consider papers that optimize cloud gaming platforms from specific aspects; usually each work focuses on optimizing one or a few components. Such studies can be further categorized into two groups: cloud server infrastructure and communications.
    a) **Cloud Server Infrastructure (Section IV-A):** The existing studies on optimizing cloud server infrastructure are surveyed. Several papers study the **Resource Allocation** problem of server and network resources among multiple data centers, server nodes, and game clients to optimize the overall cloud gaming experience, where diverse criteria are considered. Other papers optimize the **Distributed Architectures** of cloud gaming platforms, e.g., using Peer-to-Peer (P2P) overlays or multi-tier clouds for better performance and scalability.
    b) **Communications (Section IV-B):** We survey the existing work on optimizing the efficiency of content streaming over the dynamic and heterogeneous communication channels. These studies are further classified into two groups. First, several papers consider the problem of **Data Compression**, e.g., layered coding and graphics

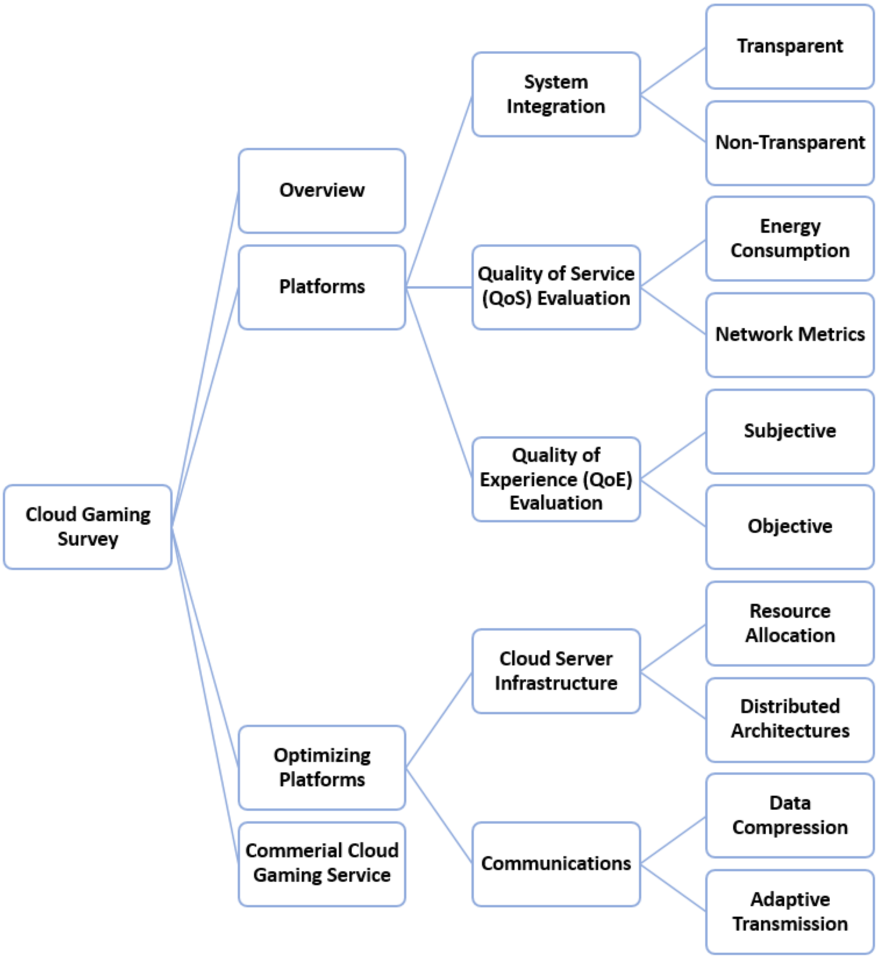

FIGURE 2. Our proposed classification system of cloud gaming papers.

compression are proposed, which may outperform the conventional 2D image compression in certain environments. Second, there are papers on **Adaptive Transmission**, which cope with the network dynamics by continuously changing various parameters, such as encoding bitrate, frame rate, and image resolution. The same adaptive transmissions may also be used to absorb the negative impacts due to insufficient resources on cloud servers and game clients.

4) **Commercial Cloud Gaming Services (Section V):** We survey the representative commercial cloud gaming services, and classify them along different aspects. We also discuss the advantages and disadvantage of different cloud gaming services.

In the rest of this article, we survey the four classes of papers in Sections II–V. They are followed by Section VI, which concludes the survey.

II. CLOUD GAMING OVERVIEW PAPERS

As a promising cloud service provisioning paradigm, cloud gaming has attracted interests from prominent research teams all over the world. These teams have shared their thoughts and ideas on cloud gaming from their viewpoints in several high-level overview papers. In this section, we survey and summarize the representative papers along this direction. Our concise summary puts readers into the context of cloud gaming research, while interested readers may find new research directions in the surveyed overview papers.

Ross [74] is the first literature that introduces the cloud gaming model to the academia in 2009, nine years after the G-cluster's demonstration of cloud gaming technology at E3. The author describes gaming as cloud computing's killer app and depicts the blueprint of novel gaming delivery paradigm, proposed by Advanced Micro Devices (AMD), which renders games' scene videos, compresses them, and transmit them to the gamers through the Internet. This approach enables online gamers to offload their graphic rendering tasks to the cloud, thus, eliminates the computational workload on gamers' local platforms. This is the most popular definition of cloud gaming adopted by most of the research work in this area. However, a recent publication [59] provides a more general definition, by envisioning the cloud gaming system as a novel computer architecture that leverages cloud resources to improve gaming performance, such as rendering, response time, precision and fairness. The authors distribute system workload to multiple cloud servers and game clients to enable this vision. For a further step, Cai et al. [4] explore the essence of cloud games as inter-dependent components, thus, define cloud gaming as utilizing cloud resources to host gaming components, thereby reducing workload at gamers' local platforms and increasing the overall system performance. According to different integration approach of the cloud, the authors identify and discuss the research directions of three cloud gaming architectures, which are *Remote Rendering*, *Local Rendering*, and *Cognitive Resource Allocation*.

After the official launch of OnLive in March 2010, the business model for cloud gaming becomes a hot topic in research society. Riungu-kalliosaari et al. [73] conduct interviews in small and medium size gaming companies to qualitatively study the adoption dynamics of cloud computing. With grounded theory method, the authors observe that the concept of cloud gaming are relatively well-known in the industry, while gaming organizations still hesitate in adopting cloud computing services and technologies due to the lack of clear business models and success stories. To this end, Ojala and Tyrvainen [66] start their investigations on developing business models for cloud gaming services. As a case study for Software as a Service (SaaS), the authors select G-cluster, one of famous cloud gaming companies, and study its business model over five years from 2005 to 2010. They conclude that, over time, the business model in cloud gaming becomes simpler and has fewer actors, which increases the revenue per gamer. In addition, they also expect the cloud gaming solution will make illegal copying practically impossible. Another work [61] considers the convergence of mobile cloud in gaming industry from a business model proposition. The authors discuss the first sketch of a possible business model of Kusanagi project, a proposed end-to-end infrastructure, from domains of service, technology, organization, and financial, while compare these domains of three cloud examples, i.e., G-cluster, Gaikai, and OnLive.

During the decade of development, there have been cloud gaming systems and services in the market. A number of positioning papers consider these systems and envision the opportunities, challenges, and directions in this area. The literature, e.g., [4], [11], [17], [23], [59], [85], [100], covers both the commercial and academic platforms, while their concerns of open issues are greatly overlapped in the topics of response time minimization, graphical video encoding, network aware adaption, QoE optimization, and cloud resource management.

Besides of these common focuses, each research team has particular interests and directions. Dey [23] concentrate on developing device aware scalable applications, which involve open issues of extending cloud to wireless networks. Soliman et al. [85] briefly discuss related legal issues, including patents, ownership concerns, guaranteed service levels, and pricing schemes. In contrast, piracy and hacking may no long be an issue, since the executable game program will not be delivered to the gamers. Wu [100] explores cloud gaming architecture from the aspect of cloud computing's three layers, i.e., IaaS, SaaS and PaaS. The author identifies security as a potential challenge in cloud gaming, especially data protection and location. Cai et al. [4] investigate the features of different game genres and identify their impact on cloud gaming system design. In addition, they provide a vision on GaaS provisioning for mobile devices. Mishra et al. [59] explain how to enhance the quality of online gaming by integrating techniques from cloud gaming research communities. Featured topics include the interplay between QoS and QoE metrics, game models, and cloud expansion. Chen et al. [11] point out some unique research directions in cloud gaming, such as game integration, visualization, user interface, server selection, and resource scheduling. Chuan et al. [17] study cloud gaming from a green media perspective. They discuss the major cloud gaming subsystems with green designs, which include a cloud data centre, graphics rendering modules, video compression techniques, and network delivery methods.

In addition to these high-level studies, more cloud gaming papers focus on individual research problems. We divide them into several classifications and survey them in the following sections.

III. CLOUD GAMING PLATFORMS

This section presents the work related to cloud gaming platforms in three steps: (i) integrated cloud gaming platforms for complete prototype systems, (ii) measurement studies on QoS metrics, and (iii) measurement studies on QoE metrics.

A. SYSTEM INTEGRATION

Providing an easy-to-use platform for (cloud) game developers is very challenging. This is because of the complex, distributed, and heterogeneous nature of the cloud gaming platforms. In fact, there is a clear tradeoff between *development complexity* and *optimization room*. Platforms opt for very low (or even no) additional development complexity may suffer from limited room for optimization, which are referred to as *transparent platforms* that run unmodified games. In contrast, other platforms opt for more optimized performance at the expense of requiring additional development complexity, such as code augmentation and recompilation, which are called *non-transparent platforms*. These two classes of cloud gaming platforms have advantages and disadvantages, and we describe representative studies in individual classes below.

The transparent platforms ease the burden of deploying new games on cloud gaming platforms, at the expense of potentially suboptimal performance. Depasquale et al. [22] present a cloud gaming platform based on the RemoteFX extension of Windows remote desktop protocol. Modern Windows servers leverage GPUs and Hyper-V virtual machines to enable various remote applications, including cloud games. Their experiments reveal that RemoteFX allows Windows servers to better adapt to network dynamics, but still suffers from high frame loss rate and inferior responsiveness. Kim et al. [44] propose another cloud gaming platform, which consists of a distributed service platform, a distributed rendering system, and an encoding/streaming system. Their platform supports isolated audio/video capturing, multiple clients, and browser-based clients. Real experiments with 40 subjects have been done, showing high responsiveness. Both Depasquale et al. [22] and Kim et al. [44] are proprietary platforms, and are less suitable for cloud gaming research. GamingAnywhere [38], [40] is the first open source transparent cloud gaming platform. Its design principles can be summarized as extensive, portable, configurable, and open. The GamingAnywhere server supports Windows and Linux, and the GamingAnywhere client runs on Windows, Linux, Mac OS, and Android. It is shown that GamingAnywhere outperforms several commercial/proprietary cloud gaming platforms, and has been used and enhanced in several cloud gaming studies in the literature. For example, Hong et al. [35] develop adaptation algorithms for multiple gamers, to maximize the gamer experience. In addition to: (i) a user study to map cloud gaming parameters to gamer experience and (ii) optimization algorithms for resource allocation, they also enhance GamingAnywhere [38], [40] to support on-the-fly adaption of frame rate and bitrate.

The non-transparent platforms require augmenting and recompiling existing games to leverage unique features for better gaming experience, which may potentially be time-consuming, expensive, and error-prone. For example, current games can be ported to Google's Native Client technology [62], [63] or to Mozilla's *asm.js* language [1], [24]. Several other studies focus on integrating new techniques with cloud gaming platforms for better gaming experience. Nan et al. [64] propose a joint video and graphics streaming system for higher coding efficiency as well. Moreover, they present a rate adaptation algorithm to further minimize the bandwidth consumption. Lee et al. [48], [49] present a system to improve the responsiveness of mobile cloud gaming by compensating network delay. In particular, their system pre-renders potential future frames based on some prediction algorithm and delivers the rendered frames to mobile clients when the network conditions are good. These frames are then used to compensate late video frames due to unstable networks. They integrate the proposed system with two open source games, and conduct a user study of 23 subjects. The subjects report good gaming experience under nontrivial network delay, as high as 250 ms. Cai et al. [3] build a prototype platform for decomposed cloud gaming, and rigorously address several system issues, which were not thoroughly investigated in their earlier work [4]. Their main contribution is the very first cognitive cloud gaming platform that automatically adapts to distributive workload in run-time, in order to optimally utilize distributed resources (on different entities, like cloud servers, in-network computing nodes, and gamers' local platforms) for the best gamer experience. On the resulting platform, several games are developed and empirically evaluated, demonstrating the potentials of cognitive cloud gaming platforms. Several enhancements on such a platform are still possible, such as implementing more sophisticated games, supporting more gamers, and providing more completed SDK (Software Development Kit) to cloud game developers.

B. QUALITY OF SERVICE EVALUATIONS

Performing QoS measurements is crucial for quantifying the performance of the cloud gaming platforms. Moreover, doing so in real-time allows us to effectively troubleshoot and even to dynamically optimize the cloud gaming platforms. The QoS related cloud gaming papers are roughly categorized into two classes: (i) energy consumption and (ii) network metrics. They are surveyed in the following.

1) ENERGY CONSUMPTION

Games have been known to push consumer computing platforms to their maximum capacity. In traditional systems such as desktop computers, it is often expected and accepted that Game software will push a system to its limits. However, mobile environments are in a strikingly different scenario as they have limited power reserves. A fully utilized mobile device may have a greatly reduced running time, thus it is important to reduce the complexity of these game software for mobile devices. Luckily, cloud gaming systems provide a potential way forward by offloading complicated processing tasks such as 3D rendering and physics calculations to powerful cloud servers. However, cares must be taken because the decoding of video, especially high definition video is far from a trivial task. We will cover some pioneering work [29], [39], [91] that has been done on this important subject.

Hans et al. [29] systematically test the energy performance of their in-house cloud gaming server *MCGS.KOM* on real world tablets. They find that when WLAN was used as the access network, cloud game software could save between 12% and 38% of energy use, depending on the types of games and tablets. Explorations on important energy saving coding parameters for H.264/AVC are reported in Taher et al. [91]. Further, Huang et al. [39] explore the energy consumption of the cloud gaming video decoders. The researchers found that frame rate has the largest impact on the decoders energy consumption, with bit rate and resolution also being major contributors. Moreover, Shea et al. [79] explore the performance and energy implications of combing cloud gaming systems with live broadcasting systems such as Twitch.

2) NETWORK METRICS

Like many other distributed multimedia applications, user experience highly depends on network conditions. Therefore, evaluating different network metrics in cloud gaming is crucial, and we present detailed survey below.

Claypool [18] measures the contents variety of different game genres in details. 28 games from 4 perspectives, including *First-Person Linear*, *Third-Person Linear*, *Third-Person Isometric*, and *Omnipresent*, are selected to analyze their scene complexity and motion, indicated by average

Intra-coded Block Size (IBS) and Percentage of Forward/backward or Intra-coded Macroblocks (PFIM), respectively. Measurements conducted by the author suggest that Microsoft's remote desktop achieves better bitrate than NoMachine's NX client, while NX client has higher frame rate. A following work [21] investigates OnLive's network characteristics, such as the data size and frequency being sent and the overall downlink and uplink bitrates. The authors reveal that the high downlink bitrates of OnLive games are very similar to those of live videos, nevertheless, OnLive's uplink bitrates are much more moderate, which are comparable to traditional game uplink traffic. They also indicate that the game traffic features are similar for three types of game genres, including *First-Person*, *Third-Person*, and *Omnipresent*, while the total bitrates can vary by as much as 50%. Another important finding is that OnLive does not demonstrate its ability in adapting bitrate and frame rates to network latency.

Chen et al. [10] analyze a cloud gaming system's response delays and segment it into three components, including network delay, processing delay, and playout delay. With this decomposition, the authors propose a methodology to measure the latency components and apply the methodology on OnLive and StreamMyGame, two of the popular cloud gaming platforms. The authors identify that OnLive system outperforming StreamMyGame in terms of latency, due to the different resource provisioning strategy based on game genres. A following work [9] by the same group extend the model by adding game delay, which represents the latency introduced by the game program to process commands and render the next video frame of the game scene. They also study how system design and selective parameters affect responsiveness, including scene complexity, updated region sizes, screen resolutions, and computation power. Their observation in network traffics are inline with previous work conducted by Claypool et al. [21]. Lower network quality, including the higher packet loss rate and insufficient bandwidth, will impose negative impacts on both of OnLive and StreamMyGame, resulting lower frame rates and worse graphic quality. Moreover, by quantifying the streaming quality, the authors further reveal that OnLive implements an algorithm to adapt its frame rate to the network delay, while StreamMyGame doesn't.

Manzano et al. [55] collect and compare network traffic traces of OnLive and Gaikai, including packet inter-arrival times, packet size, and packet inter-departure time, to observe the difference between cloud gaming and traditional online gaming from the perspectives of network load and traffic characteristics. The authors reveal that the package size distributions between the two platforms are similar, while the packet inter-arrival times are distinct. Afterwards, Manzano et al. [56] claim to be the first research work on specific network protocols used by cloud gaming platforms. They focus on conducting a reverse engineering study on OnLive, based on extensive traffic traces of several games. The authors further propose a per-flow traffic model for OnLive, which can be used for network dimensioning, planning optimization, and other studies.

Shea et al. [81] measure the interaction delay and image quality of OnLive system, under diverse games, computers, and network configurations. The authors conclude that cloud procedure introduces 100 to 120 ms latency to the overall system, which requires further developments in both video encoders and streaming software. Meanwhile, the impacts of compression mechanism on video quality are quite noticeable, especially under the circumstances with lower available bandwidth. They later present an experimental study [80] on the performance of existing commercial games and ray-tracing applications with graphical processing units (GPUs). According to their analysis, gaming applications in virtualized environments demonstrate poorer performance than the instances executing in non-virtualized bare-metal baseline. Detailed hardware profiling further reveals that the pass-through access introduces memory bottleneck, especially for those games with real-time interactions. Another work [36], however, observes more advanced virtualization technologies such as mediated pass-through maintain high performance in virtualized environments. In the authors' measurement work, rendering with virtualized GPUs may achieves better performance than direct pass-through ones. In addition, if the system adopts software video coding, the CPU may became the bottleneck, while hypervisor will no longer be the constraint of the system performance. Based on these analysis, the authors conclude that current virtualization techniques are already good enough for cloud gaming.

Suznjevic et al. [89] measure 18 games on GamingAnywhere [38] to analyze the correlation between the characteristics of the games played and their network traffic. The authors observe the highest values for motion, action game and shooter games, while the majority of strategy games are relatively low. In contrast, for spatial metrics the situation is reversed. They also conclude that the bandwidth usage for most games are within the range of 3 and 4 Mbit/s, except the strategy games that consume less network resources. Another notable finding is that, gamers' action rate will introduce a slight packet rate increase, but will not affect the generated network traffic volume.

Lampe et al. [46] conduct experimental evaluations of user-perceived latency in cloud games and locally executed video games. Their results, produced by a semi-automatic measurement tool called GALAMETO.KOM, indicate that cloud gaming introduces additional latency to game programs, which is approximately 85% to 800% higher than local executions. This work also features the significant impact of round-trip time. The measurement results confirm the hypothesis that the geographical placement of cloud data centres is an important element in determining response delay, specifically when the cloud gaming services are accessed through cellular networks.

Xue et al. [102] conduct a passive and active measurement study for CloudUnion, a Chinese cloud gaming system. The authors characterize the platform from the aspects of architecture, traffic pattern, user behaviour, frame rate and gaming latency. Observations include: (i) CloudUnion adopts a geo-distributed infrastructure; (ii) CloudUnion suffers from a queuing problem with different locations from time to time; (iii) the User Datagram Protocol (UDP) outperforms the Transmission Control Protocol (TCP) in terms of response delay while sacrificing the video quality; and (iv) CloudUnion adopts conservative video rate recommendation strategy. By comparing CloudUnion and GamingAnywhere [38], the authors observe four common problems. First, the uplink and downlink data rates are asymmetric. Second, low-motion games perceive a periodical jitter at the interval of 10 seconds. Third, audio and video streams are suffering from synchronization problem. Fourth, packet loss in network transmission degrades gaming experiences significantly.

C. QUALITY OF EXPERIENCE EVALUATIONS

Measuring and modeling cloud gaming QoE are no easy tasks because QoE metrics are subjective. In particular, enough subjects need to be recruited, and time-consuming, tedious, and expensive user studies need to be carried out. After that, practical models to relate the QoS and QoE metrics need to be proposed, trained, and evaluated. Only when the resulting models are validated with large datasets, they can be employed in actual cloud gaming platforms. Cloud gaming QoE has been studied in the literature and can be categorized into two classes: (i) general cloud gaming QoE evaluations, and (ii) mobile cloud gaming QoE evaluations, which are tailored for mobile cloud games, where mobile devices are resource constrained and vulnerable to inferior wireless network conditions. We survey the related work in these two classes below.

Chang et al. [8] present a measurement and modeling methodology on cloud gaming QoE using three popular remote desktop systems. Their experiment results reveal that the QoE (in gamer performance) is a function of frame rate and graphics quality, and the actual functions are derived using regression. They also show that different remote desktop systems lead to quite diverse QoE levels under the same network conditions. Jarschel et al. [42] present a testbed for a user study on cloud gaming services. Mean Opinion Score (MOS) values are used as the QoE metrics, and the resulting MOS values are found to depend on QoS parameters, such as network delay and packet loss, and context, such as game genres and gamer skills. Their survey also indicates that very few gamers are willing to commit themselves in a monthly fee plan for cloud gaming. Hence, better business models are critical to long-term success of cloud gaming. Möller et al. [60] also conduct a subjective test in the labs, and consider 7 different MOS values: input sensitivity, video quality, audio quality, overall quality, complexity, pleasantness, and perceived value. They observe complex interplays among QoE metrics, QoS metrics, testbed setup, and software implementation. For example, the rate control algorithm implemented in cloud gaming client is found to interfere with the bandwidth throttled by a traffic shaper. Several open issues are raised after analyzing the results of the user study, partially due to the limited number of participants. Slivar et al. [84] carry out a user study of in-home cloud gaming, i.e., the cloud gaming servers and clients are connected over a LAN. Several insights are revealed, e.g., switching from a standard game client to in-home cloud gaming client leads to QoE degradation, measured in MOS values. Moreover, more skilled gamers are less satisfied with in-home cloud gaming. Hossain et al. [37] adopt gamer emotion as a QoE metric and study how several screen effects affect gamer emotion. Sample screen effects include adjusting: (i) redness, (ii) blueness, (iii) greenness, (iv) brightness, and (v) contrast; and the goal of applying these screen effects is to mitigate negative gamer emotion. They then perform QoE optimization after deriving an empirical model between screen effects and gamer emotion.

Some other QoE studies focus on the response delay, which is probably the most crucial performance metric in cloud gaming, where servers may be geographically far away from clients. Lee et al. [50] find that response delay imposes different levels of implications on QoE with different game genres. They also develop a model to capture this implication as a function of gamer inputs and game scene dynamics. Quax et al. [71] make similar conclusions after conducting extensive experiments, e.g., gamers playing action games are more sensitive to high responsive delay. Claypool and Finkel [20] perform user studies to understand the objective and subjective effects of network latency on cloud gaming. They find that both MOS values and gamer performance degrade linearly with network latency. Moreover, cloud gaming is very sensitive to network latency, similar to the traditional first-person avatar games. Raaen [72] designs a user study to quantify the smallest response delay that can be detected by gamers. It is observed that some gamers can perceive $<$ 40 ms response delay, and half of the gamers cannot tolerate $\geq$ 100 ms response delay.

Huang et al. [41] perform extensive cloud gaming experiments using both mobile and desktop clients. Their work reveals several interesting insights. For example, gamers' satisfaction on mobile clients are more related to graphics quality, while the case on desktop clients is more correlated to control quality. Furthermore, graphics and smoothness quality are significantly affected by the bitrate, frame rate, and network latency, while the control quality is determined only by the client types (mobile or desktop). Wang and Dey [94], [97] build a mobile cloud gaming testbed in their lab for subjective tests. They propose a Game Mean Opinion Score (GMOS) model, which is a function of game genre, streaming configuration, measured Peak Signal-to-Noise Ratio (PSNR), network latency, and packet loss. The derivations of model parameters are done via offline regression, and the resulting models can be used for optimizing mobile cloud gaming experience. Along this line, Liu et al. [54] propose a Cloud Mobile Rendering–Mean Opinion Score (CMR-MOS) model, which is a variation of GMOS. CMR-MOS has been used in selecting detail levels of remote rendering applications, like cloud games.

IV. OPTIMIZING CLOUD GAMING PLATFORMS

This section surveys optimization studies on cloud gaming platforms, which are further divided into two classes: (i) cloud server infrastructure and (ii) communications.

A. CLOUD SERVER INFRASTRUCTURE

To cope with the staggering demands from the massive number of cloud gaming users, carefully-designed cloud server infrastructures are required for high-quality, robust, and sustainable cloud gaming services. Cloud server infrastructures can be optimized by: (i) intelligently allocating resources among servers or (ii) creating innovative distributed structures. We detail these two types of work in the following.

1) RESOURCE ALLOCATION

The amount of resources allocated to high performance multimedia applications such as cloud gaming continues to grow in both public and private data centers. The high demand and utilization patterns of these platforms make the smart allocation of these resources paramount to the efficiency of both public and private clouds. From Virtual Machine (VM) placement to shared GPUs, researchers from many areas have been exploring how to efficiently use the cloud to host cloud gaming platforms. We now explore the important work done in this area to facilitate efficient deployment of cloud gaming platforms.

Critical work has been done on both VM placement and cloud scheduling to facilitate better quality of cloud gaming services. For example, Wang et al. [98] show that, with proper scheduling of cloud instances, cloud gaming servers could be made wireless networking aware. Simulations of their proposed scheduler show the potential of increased performance and decreased costs for cloud gaming platforms. Researchers also explore making resource provisioning cloud gaming aware. For example, a novel QoE aware VM placement strategy for cloud gaming is developed [33]. Further, research has been done to increase the efficiency of resource provisioning for massively multi-player online games (MMOG) [57]. The researchers develop greedy heuristics to allocate the minimum number of computing nodes required to meet the MMOG service needs. Researchers also study the popularity of games on the cloud gaming service OnLive and propose methods to improve performance of these systems based on game popularity [25]. Later, a resource allocation strategy [51] based on the expected ending time of each play session is proposed. The strategy can reduce the cost of operation to cloud gaming providers by reducing the number of purchased nodes required to meet their clients needs. They note that classical placement algorithms such as *First Fit* and *Best Fit*, are not effective for cloud gaming. After extensive experiments, the authors show an algorithm leveraging on neural-network-based predictions, which could improve VM deployment, and potentially decreases operating costs.

Although many cloud computing workloads do not require a dedicated GPU, cloud gaming servers require access to a rendering device to provide 3D graphics. As such VM and workload placements have been researched to ensure cloud gaming servers have access to adequate GPU resources. Kim et al. [45] propose a novel architecture to support multiple-view cloud gaming servers, which share a single GPU. This architecture provides multi-focal points inside a shared cloud game, allowing multiple gamers to potentially share a game world, which is rendered on a single GPU. Zhao et al. [104] perform an analysis of the performance of combined CPU/GPU servers for game cloud deployments. They try offloading different aspects of game processing to these cloud servers, while maintaining some local processing at the client side. They conclude that keeping some processing at the client side may lead to an increase in QoS of cloud gaming platforms.

Pioneering research has also been done on GPU sharing and resource isolation for cloud gaming servers [70], [103]. These works show that with proper scheduling and allocation of resources we can maximize GPUs utilization, while maintaining high performance for the gamers sharing a single GPU. Shea and Liu [80] show that direct GPU assignment to a virtualized gaming instance can lead to frame rate degradation of over 50% in some gaming applications. They find that the GPU device pass-through severely diminishes the data transfer rate between the main memory and the GPU. Their follow-up work using more advanced platforms [78] reveals that although the memory transfer degradation still exists, it no longer affects the frame rate of current generation games. Hong et al. [34] perform a parallel work, where they discover that the frame rate issue presents in virtualized clouds may be mitigated by using mediated pass-through, instead of direct assignment.

In addition, work has been done to augment existing clouds and games to improve cloud gaming efficiency. It has been shown that using game engine information can greatly reduce the resources needed to calculate the motion estimation (ME) needed for conventional compression algorithms such as H.264/AVC [76]. Research into these technique shows that we can accelerate the motion estimation phase by over 14% if we use in-game information for encoding. Others have proposed using reusable modules for cloud gaming servers [30]. They refer to these reusable modules as substrates and test the latency between the different components. All these data compression studies affect resource allocation; we provide a comprehensive survey on data compression for cloud gaming in Section IV-B.1.

2) DISTRIBUTED ARCHITECTURES

Due to the vast geographic distribution of the cloud gaming clients the design of distributed architectures is of critical importance to the deployment of cloud gaming systems. The design of these systems must be carefully optimized to ensure that a cloud gaming system can sufficiently cover its target audience. Further, to maintain the extremely low delay tolerance required for high QoE even the placement of different server components must be optimized for the lowest possible latency. These innovative distributed architectures have been investigated in the literature, and we detail them below.

Sselbeck et al. [90] discover that running a cloud gaming based massively multi-player online game (MMOG) may suffer from increased latency. These issues are aggravated in a cloud gaming context because MMOG are already extremely latency sensitive applications. The increased latency introduced by a cloud gaming may vastly decrease the playability of these games. To deal with this increased latency, they propose a P2P based solution. Similarly, Prabu and Purushotham [69] propose a P2P system based on Windows Azure to support online games.

Research has also been done on issues created by the geographical distance between the end user of cloud gaming and a cloud gaming data center. Choy et al. [13] show that the current geographical deployments of public data centers leave a large fraction of the USA with an unacceptable RTT for low latency applications such as cloud gaming. To help mitigate this issue, they propose deploying edge servers near some users for cloud gaming; a follow up work further explores this architecture and shows that hybrid edge-cloud architectures could indeed expand the reach of cloud gaming data centers [14]. Similarly, Siekkinen et al. [83] propose a distributed cloud gaming architecture with servers deployed near local gamers when necessary. The researchers prototype the system and show that if being deployed widely enough, for example at the ISP level, cloud gaming could reach an even larger audience. Tian et al. [92] perform an extensive investigation into issues of deploying cloud gaming architecture with distributed data centers. They focus on a scenario where adaptive streaming technology is available to the cloud provider. The authors give an optimization algorithm, which can improve gamer QoE as well as reducing operating costs of the cloud gaming provider. The algorithm is evaluated using trace driven simulations, and the results show a potential cost savings of 25% to the cloud gaming provider.

B. COMMUNICATIONS

Due to the distributed nature of cloud gaming services, the efficiency and robustness of the communication channels between cloud gaming servers and clients are crucial and have been studied. These studies can be classified into two groups: (i) the data compression algorithms to reduce the network traffic amount and (ii) the transmission adaptation algorithms to cope with network dynamics. We survey the work in these two groups in the following.

1) DATA COMPRESSION

After game scenes are computed on cloud servers, they have to be captured in proper representations and compressed before being streamed over networks. This can be done in one of the three data compression schemes: (i) *video compression*, which encodes 2D rendered videos and potentially auxiliary videos (such as depth videos) for client side post-rendering operations, (ii) *graphics compression*, which encodes 3D structures and 2D textures, and (iii) *hybrid compression*, which combines both video and graphics compression. Upon cloud gaming servers produce compressed data streams, the servers send the streams to client computers over communication channels. We survey each of the three schemes below.

Video compression is the most widely-used data compression schemes for cloud gaming probably because 2D video codecs are quite mature. These proposals strive to improve the coding efficiency in cloud gaming, and can be further classified into groups depending on whether in-game graphics contexts, such as camera locations and orientations, are leveraged for higher coding efficiency. We first survey the proposals that do not leverage graphics contexts. Cai et al. [6] propose to cooperatively encode cloud gaming videos of different gamers in the same game session, in order to leverage inter-gamer redundancy. This is based on an observation that game scenes of close-by gamers have non-trivial overlapping areas, and thus adding inter-gamer predictive video frames may improve the coding efficiency. The high-level idea is similar to multiview video codecs, such as H.264/MVC, and the video packets shared by multiple gamers are exchanged over an auxiliary short-range ad-hoc network in a P2P fashion. Cai et al. [5] improve upon the earlier work [6] by addressing three more research problems: (i) uncertainty due to mobility, (ii) diversity of network conditions, and (iii) model of QoE. These problems are solved by a suite of optimization algorithms proposed in their work. Sun and Wu [88] solve the video rate control problem in cloud gaming in two steps. First, they adopt the concept of RoI, and define heterogeneous importance weights for different regions of game scenes. Next, they propose a macroblock-level rate control scheme to optimize the RoI-weighted video quality. Cheung et al. [12] propose to concatenate the graphic renderer with a customized video coder on servers in cellular networks and multicast the coded video stream to a gamer and multiple observers. Their key innovation is to leverage the depth information used in 3D rendering process to locate the RoI and then allocate more bits to that region. The resulting video coder is customized for cloud gaming, yet produces standard compliant video streams for mobile devices. Liu et al. [53] also leverage rendering information to improve video encoding in cloud gaming for better perceived video quality and shorter encoding time. In particular, they first analyze the rendering information to identify RoI and allocate more bits on more important regions, which leads to better perceived video quality. In addition, they use this information to accelerate the encoding process, especially the time used in motion estimation and macroblock mode selection. Experiments reveal that their proposed video coder saves 42% of encoding time and achieves perceived video quality similar to the unmodified video coder. Similarly, Semsarzadeh et al. [76] study the feasibility of using rendering information to accelerate the computationally-intensive motion estimation and demonstrate that it is possible to save 14.32% of the motion estimation time and 8.86% of the total encoding time. The same authors [77] then concertize and enhance their proposed method, in which they present the general method, well-designed programming interface, and detailed motion estimation optimization. Both subjective and objective tests show that their method suffers from very little quality drop compared to the unmodified video coder. It is reported that they achieve 24% and 39% speedups on the whole encoding process and motion estimation, respectively.

Next, we survey the proposals that utilize graphics contexts [82], [101]. Shi et al. [82] propose a video compression scheme for cloud gaming, which consists of two unique techniques: (i) 3D warping-assisted coding and (ii) dynamic auxiliary frames. 3D warping is a light-weight 2D post-rendering process, which takes one or multiple *reference* view (with image and depth videos) to generate a *virtual* view at a different camera location/orientation. Using 3D warping allows video coders to skip some video frames, which are then wrapped at client computers. Dynamic auxiliary frames refer to those video frames rendered with intelligently-chosen camera location/orientations that are not part of the game plays. They show that the auxiliary frames help to improve 3D warping performance. Xu et al. [101] also propose two techniques to improve the coding efficiency in cloud gaming. First, the camera rotation is rectified to produce video frames that are more motion estimation friendly. On client computers, the rectified videos are compensated with some camera parameters using a light-weight 2D process. Second, a new interpolation algorithm is designed to preserve sharp edges, which are common in-game scenes. Last, we notice that the video compression schemes are mostly orthogonal to the underneath video coding standards, and can be readily integrated with the recent (or future) video codecs for further performance improvement.

Graphics compression is proposed for better scalability, because 3D rendering is done at individual client computers. Compressing graphics data, however, is quite challenging and may consume excessive network bandwidth [52], [58]. Lin et al. [52] design a cloud gaming platform based on graphics compression. Their platform has three graphics compression tools: (i) intra-frame compression, (ii) inter-frame compression, and (iii) caching. These tools are applied to graphics commands, 3D structures, and 2D textures. Meilnder et al. [58] also develop a similar platform for mobile devices, where the graphics are sent from cloud servers to proxy clients, which then render game scenes for mobile devices. They also propose three graphics compression tools: (i) caching, (ii) lossy compression, and (iii) multi-layer compression. Generally speaking, tuning cloud gaming platforms based on graphics compression for heterogeneous client computers is non-trivial, because mobile (or even some stationary) computers may not have enough computational power to locally render game scenes.

Hybrid compression [15], [16] attempts to fully utilize the available computational power on client computers to maximize the coding efficiency. For example, Chuah and Cheung [15] propose to apply graphics compression on simplified 3D structures and 2D textures, and send them to client computers. The simplified scenes are then rendered on client computers, which is called the *base* layer. Both the full-quality video and the base-layer video are rendered on cloud servers, and the residue video is compressed using video compression and sent to client computers. This is called the *enhancement* layer. Since the base layer is compressed as graphics and the enhancement layer is compressed as videos, the proposed approach is a hybrid scheme. Based on the layered coding proposal, Chuah et al. [16] further propose a complexity-scalable base-layer rendering pipeline suitable for heterogeneous mobile receivers. In particular, they employ scalable Blinn-Phong lighting for rendering the base-layer, which achieves maximum bandwidth saving under the computing constraints of mobile receivers. Their experiments demonstrate that their hybrid compression solution, customized for cloud gaming, outperforms single-layer general-purpose video codecs.

2) ADAPTIVE TRANSMISSION

Even though data compression techniques have been applied to reduce the network transmission rate, the fluctuating network provisioning still results in unstable service quality to the gamers in cloud gaming system. These unpredictable factors include bandwidth, round-trip time, jitter, and etc. Under this circumstance, adaptive transmission is introduced to further optimize gamers' QoE. The foundation of these studies is based on a common sense: gamers would prefer to scarify video quality to gain smoother playing experience in insufficient network QoS supplement.

Jarvinen et al. [43] explore the approach to adapt the gaming video transmission to available bandwidth. This is accomplished by integrating a video adaptation module into the system, which estimates the network status from network monitor in real-time and dynamically manipulates the encoding parameters, such as frame rate and quantization, to produce specific adaptive bit rate video stream. The authors utilize RTT jitter value to detect the network congestion, in order to decide if the bit rate adaptation should be triggered. To evaluate this proposal, a following work [47] conducts experiments on a normal television with an IPTV set-top-box. The authors simulate the network scenarios in homes and hotels to verify that the proposed adaptation performed notably better.

Adaptive transmission has also been studied in mobile scenarios. Wang and Dey [95] first decompose the cloud gaming system's response time into sub-components: server delay, network uplink/downlink delay, and client delay. Among the optimization techniques applied, rate-selection algorithm provides a dynamic solution that determine the time and the way to switch the bit rate according to the network delay. As a further step, Wang and Dey [96] study the potential of rendering adaptation. They identify the rendering parameters that affect a particular game, including realistic effect (e.g., colour depth, multi-sample, texture-filter, and lighting mode), texture detail, view distance and enabling grass. Afterwards, they analyze these parameters' characteristics of communications and computation costs and propose their rendering adaptation scheme, which is consisted of optimal adaptive rendering settings and level-selection algorithm. With the experiments conducted on commercial wireless networks, the authors demonstrate that acceptable mobile gaming user experience can be ensured by their rendering adaption technique. Thus, they claim that their proposal is able to facilitate cloud gaming over mobile networks.

Other aspects of transmission adaptation have also been investigated in the literature. He et al. [31] consider the adaptive transmission from the perspective of multi-player. The authors calculate the packet urgency based on buffer status estimation and propose a scheduling algorithm. In addition, they also suggest an adaptive video segment request scheme, which estimates media access control (MAC) queue as an additional information to determine the request time interval for each gamer, on the purpose of improving the playback experience. Bujari et al. [2] provides a VoAP algorithm to address the flow coexistence issue in wireless cloud gaming service delivery. This research problem is introduced by the concurrent transmissions of TCP-based and UDP-based streams in home scenario, where the downlink requirement of gaming video exacerbate the operation of above mentioned transport protocols. The authors' solution is to dynamically modify the advertised window, in such way the system can limit the growth of the TCP flow's sending rate. Wu et al. [99] present a novel transmission scheduling framework dubbed AdaPtive HFR vIdeo Streaming (APHIS) to address the issue in the cloud gaming video delivery through wireless networks. The authors first propose an online video frame selection algorithm to minimize the total distortion based on network status, input video data, and delay constraint. Afterwards, they introduce an unequal forward error correction (FEC) coding scheme to provide differentiated protection for Intra (I) and Predicted (P) frames with low-latency cost. The proposed APHIS framework is able to appropriately filter video frames and adjust data protection levels to optimize the quality of HFR video streaming. Hemmati et al. [32] propose an object selection algorithm to provide an adaptive scene rendering solution. The basic idea is to exclude less important objects from the final output, thus to reduce less processing time for the server to render and encode the frames. In such a way, the cloud gaming system is able to achieve a lower bit rate to stream the resulting video. The proposed algorithm evaluates the importance of objects from the game scene based on the analysis of gamers' activities and do the selection work. Experiments demonstrate that this approach reduces streaming bit rate by up to 8.8%.

V. COMMERCIAL CLOUD GAMING SERVICES

In addition to the technical problems discussed in prior sections, commercialization and business models of cloud gaming services are critical to their success. We survey the commercialization efforts starting from a short history on cloud gaming services. G-cluster [26] starts building cloud gaming services since early 2000's. In particular, G-cluster publicly demonstrated live game streaming[1] over WiFi to a PDA in 2001, and a commercial game-on-demand service in 2004. G-cluster's service is tightly coupled with several third-party companies, including game developers, network operators, and game portals. This can be partially attributed to the less mature Internet connectivity and data centers, which force G-cluster to rely on network QoS supports from network operators. Ojala and Tyrvainen [66] presents the evolution of G-cluster's business model, and observe that the number of G-cluster's third-party companies is reduced over years. The number of households having access to G-cluster's IPTV-based cloud gaming service increased from 15,000 to 3,000,000 between 2005 and 2010.

In late 2000's, emerging cloud computing companies start offering Over-The-Top (OTT) cloud gaming services, represented by OnLive [67], Gaikai [27], and GameNow [28]. OTT refers to delivering multimedia content over the Internet above arbitrary network operators to end users, which trades QoS supports for ubiquitous access to cloud games. OnLive [67] was made public in 2009, and was a well-known cloud gaming service, probably because of its investors including Warner Bros, AT&T, Ubisoft, and Atrari. OnLive provided subscription based service, and hosted its servers in several States within the US, to control the latency due to geographical distances. OnLive ran into financial difficulty in 2012, and ceased operations in 2015 after selling their patents to Sony [87]. Gaikai [27] offered cloud gaming service using a different business model. Gaikai adopt cloud gaming to allow gamers to try new games without purchasing and installing software on their own machines. At the end of each gameplay, gamers are given options to buy the game if they like it. That is, Gaikai is more like an advertisement service for game developers to boost their sales. Gaikai was acquired by Sony [86] in 2012, which leads to a new cloud gaming service from Sony, called PS Now [68] launched in 2014. PS Now allows gamers to play PlayStation games as cloud games, and adopts two charging models: per-game and monthly subscription.

The aforementioned cloud gaming services can be classified in groups from two aspects. We discuss the advantages and disadvantages of different groups in the following. First, cloud gaming services are either: (i) integrated with underlaying networks or (ii) provided as OTT services. Tighter integration provides better QoS guarantees which potentially lead to better user experience, while OTT reduces the expenses on cloud gaming services at a possible risk of unstable and worse user experience. Second, cloud gaming services adopt one of the three charging models: (i) subscription, (ii) per-game, and (iii) free to gamers. More specifically, cloud gaming users pay for services in the first two charging models, while third-party companies, which can be game developers or network

[1] At that time, the term *cloud* was not yet popular.

operators, pay for services in the third charging model. In the future, there may be innovative ways to offer cloud gaming services to general publics in a commercially-viable manner.

VI. CONCLUSION AND OUTLOOK

In this article, we grouped the existing cloud gaming research into four classifications: (i) overview, (ii) platform, (iii) optimization, and (iv) commercialization. In Section II (overview), we included papers that introducing general and specialized (such as mobile) cloud gaming. In Section III (platform), we presented the basic cloud gaming platforms that support quantitative performance measurements. More specifically, we considered: (i) QoS evaluations, such as energy consumption and network metrics, and (ii) QoE evaluations, such as gamer experience. In Section IV (optimization), we presented the two major optimization directions: (i) cloud server infrastructure, such as resource allocation and distributed architecture, and (ii) communications, such as data compression and adaptive transmission. In Section V (commercialization), we gave a brief history of cloud gaming services, followed by the design decisions made by representative commercial cloud gaming services.

Cloud gaming is not a panacea and incurs non-trivial costs to service providers. Minimizing the cost on cloud and networking resources while achieving high gamer experience requires careful optimization like the approaches explored in this survey. Without these optimizations, service provider cannot consolidate enough cloud gaming users to each physical machine. This in turn leads to much lower profits, and may drive the service provider out of business. Some early industrial pioneers such as OnLive [67] have unfortunately exited the market. More recent cloud gaming services such as PS Now [68] and GameNow [28] are better optimized and will be more competitive in the current gaming industry. As commercial cloud gaming services become financially sustainable, the new cloud gaming ecosystem will continue to expand, leading to more investments and technologies to improve these services. Much of the innovation needed to push cloud gaming to the next level may reside in creating new programing paradigms to support the unique needs of these complex systems. Most current cloud gaming platforms work as a ''black box'' simply wrapping a traditionally programed game in a support system to enable cloud gaming. Although, the original black box model of cloud gaming has led to many practical real world implementations a more integrated approach may be necessary. It is likely that using in-game contexts or whole new programing paradigms may solve some of cloud gaming shortcomings [7]. Future cloud gaming aware programing paradigms will help facilitate both better user experience and resource utilization. This will allow more innovative, yet demanding ideas to be implemented, which in turn results in critical momentum towards building the next generation cloud gaming services.

In summary, the advances of technologies turn playable cloud gaming services into reality; more optimization techniques gradually make cloud gaming services profitable; hence, we believe that we are on the edge of a new era of a whole new cloud gaming ecosystem, which will eventually leads to the next generation cloud gaming services.

REFERENCES

[1] (Mar. 2013). *asm.js*. [Online]. Available: http://asmjs.org/

[2] A. Bujari, M. Massaro, and C. Palazzi, ''Vegas over access point: Making room for thin client game systems in a wireless home,'' *IEEE Trans. Circuits Syst. Video Technol.*, vol. 25, no. 12, pp. 2002–2012, Dec. 2015.

[3] W. Cai, H. Chan, X. Wang, and V. Leung, ''Cognitive resource optimization for the decomposed cloud gaming platform,'' *IEEE Trans. Circuits Syst. Video Technol.*, vol. 25, no. 12, pp. 2038–2051, Dec. 2015.

[4] W. Cai, M. Chen, and V. C. M. Leung, ''Toward gaming as a service,'' *IEEE Internet Comput.*, vol. 18, no. 3, pp. 12–18, May/Jun. 2014.

[5] W. Cai, Z. Hong, X. Wang, H. Chan, and V. Leung, ''Quality-of-experience optimization for a cloud gaming system with ad hoc cloudlet assistance,'' *IEEE Trans. Circuits Syst. Video Technol.*, vol. 25, no. 12, pp. 2092–2104, Dec. 2015.

[6] W. Cai, V. Leung, and L. Hu, ''A cloudlet-assisted multiplayer cloud gaming system,'' *Mobile Netw. Appl.*, vol. 19, no. 2, pp. 144–152, Nov. 2013.

[7] W. Cai *et al.*, ''The future of cloud gaming,'' *Proc. IEEE*, vol. 104, no. 4, pp. 687–691, Apr. 2016.

[8] Y.-C. Chang, P.-H. Tseng, K.-T. Chen, and C.-L. Lei, ''Understanding the performance of thin-client gaming,'' in *Proc. IEEE Int. Workshop Tech. Committee Commun. Quality Rel. (CQR)*, Naples, FL, USA, May 2011, pp. 1–6.

[9] 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 Trans. Multimedia*, vol. 16, no. 2, pp. 480–495, Feb. 2014.

[10] K. Chen, Y. Chang, P. Tseng, C. Huang, and C. Lei, ''Measuring the latency of cloud gaming systems,'' in *Proc. ACM Int. Conf. Multimedia (MM)*, Scottsdale, AZ, USA, Nov. 2011, pp. 1269–1272.

[11] K.-T. Chen, C.-Y. Huang, and C.-H. Hsu, ''Cloud gaming onward: Research opportunities and outlook,'' in *Proc. IEEE Conf. Multimedia Expo Workshops (ICMEW)*, Chengdu, China, Jul. 2014, pp. 1–4.

[12] G. Cheung, T. Sakamoto, and W. Tan, ''Graphics-to-video encoding for 3G mobile game viewer multicast using depth values,'' in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Singapore, Oct. 2004, pp. 2805–2808.

[13] 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 *Proc. 11th IEEE Annu. Workshop Netw. Syst. Support Games (NetGames)*, Venice, Italy, Nov. 2012, pp. 1–6.

[14] S. Choy, B. Wong, G. Simon, and C. Rosenberg, ''A hybrid edge-cloud architecture for reducing on-demand gaming latency,'' *Multimedia Syst.*, vol. 20, no. 5, pp. 503–519, Oct. 2014.

[15] S.-P. Chuah and N.-M. Cheung, ''Layered coding for mobile cloud gaming,'' in *Proc. Int. Workshop Massively Multiuser Virtual Environ. (MMVE)*, Singapore, Mar. 2014, pp. 1–6.

[16] S.-P. Chuah, N.-M. Cheung, and C. Yuen, ''Layered coding for mobile cloud gaming using scalable Blinn–Phong lighting,'' *IEEE Trans. Image Process.*, vol. 25, no. 7, pp. 3112–3125, Jul. 2016.

[17] S.-P. Chuah, C. Yuen, and N.-M. Cheung, ''Cloud gaming: A green solution to massive multiplayer online games,'' *IEEE Wireless Commun.*, vol. 21, no. 4, pp. 78–87, Aug. 2014.

[18] M. Claypool, ''Motion and scene complexity for streaming video games,'' in *Proc. Int. Conf. Found. Digit. Games (FDG)*, Orlando, FL, USA, Apr. 2009, pp. 34–41.

[19] M. Claypool and K. Claypool, ''Latency and player actions in online games,'' *Commun. ACM*, vol. 49, no. 11, pp. 40–45, Nov. 2006.

[20] M. Claypool and D. Finkel, ''The effects of latency on player performance in cloud-based games,'' in *Proc. 13th Annu. Workshop Netw. Syst. Support Games (NetGames)*, Nagoya, Japan, Dec. 2014, pp. 1–6.

[21] M. Claypool, D. Finkel, A. Grant, and M. Solano, ''Thin to win? Network performance analysis of the OnLive thin client game system,'' in *Proc. 11th IEEE Annu. Workshop Netw. Syst. Support Games (NetGames)*, Venice, Italy, Nov. 2012, pp. 1–6.

[22] E. Depasquale *et al.*, ''An analytical method of assessment of RemoteFX as a cloud gaming platform,'' in *Proc. IEEE Conf. Medit. Electrotech. Conf. (MELECON)*, Beirut, Lebanon, Apr. 2014, pp. 127–133.

[23] S. Dey, ''Cloud mobile media: Opportunities, challenges, and directions,'' in *Proc. IEEE Conf. Comput., Netw. Commun. (ICNC)*, Maui, HI, USA, Feb. 2012, pp. 929–933.

[24] (Aug. 2013). *Emscripten*. [Online]. Available: http://emscripten.org

[25] D. Finkel, M. Claypool, S. Jaffe, T. Nguyen, and B. Stephen, "Assignment of games to servers in the OnLive cloud game system," in *Proc. Annu. Workshop Netw. Syst. Support Games (NetGames)*, Nagoya, Japan, Dec. 2014, Art. no. 4.

[26] (Jan. 2015). *G-Cluster*. [Online]. Available: http://www.gcluster.com/eng

[27] (Jan. 2015). *GaiKai*. [Online]. Available: http://www.gaikai.com/

[28] (Jan. 2015). *GameNow*. [Online]. Available: http://www.ugamenow.com

[29] R. Hans, U. Lampe, D. Burgstahler, M. Hellwig, and R. Steinmetz, "Where did my battery go? Quantifying the energy consumption of cloud gaming," in *Proc. IEEE Int. Conf. Mobile Services (MS)*, Anchorage, AK, USA, Jun. 2014, pp. 63–67.

[30] M. Hassam, N. Kara, F. Belqasmi, and R. Glitho, "Virtualized infrastructure for video game applications in cloud environments," in *Proc. ACM Symp. Mobility Manage. Wireless Access (MobiWac)*, Montreal, QC, Canada, Sep. 2014, pp. 109–114.

[31] L. He, G. Liu, and C. Yuchen, "Buffer status and content aware scheduling scheme for cloud gaming based on video streaming," in *Proc. IEEE Conf. Multimedia Expo Workshops (ICMEW)*, Chengdu, China, Jul. 2014, pp. 1–6.

[32] M. Hemmati, A. Javadtalab, A. Shirehjini, S. Shirmohammadi, and T. Arici, "Game as video: Bit rate reduction through adaptive object encoding," in *Proc. ACM Workshop Netw. Oper. Syst. Support Digit. Audio Video (NOSSDAV)*, Oslo, Norway, Feb. 2013, pp. 7–12.

[33] H.-J. Hong, D.-Y. Chen, C.-Y. Huang, K.-T. Chen, and C.-H. Hsu, "QoE-aware virtual machine placement for cloud games," in *Proc. Annu. Workshop Netw. Syst. Support Games (NetGames)*, Denver, CO, USA, Dec. 2013, pp. 1–2.

[34] H.-J. Hong, D.-Y. Chen, C.-Y. Huang, K.-T. Chen, and C.-H. Hsu, "Placing virtual machines to optimize cloud gaming experience," *IEEE Trans. Cloud Comput.*, vol. 3, no. 1, pp. 42–53, Jan. 2015.

[35] H.-J. Hong, C.-F. Hsu, T.-H. Tsai, C.-Y. Huang, K.-T. Chen, and C.-H. Hsu, "Enabling adaptive cloud gaming in an open-source cloud gaming platform," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 25, no. 12, pp. 2078–2091, Dec. 2015.

[36] H.-J. Hong, T.-Y. Fan-Chiang, C.-R. Lee, K.-T. Chen, C.-Y. Huang, and C.-H. Hsu, "GPU consolidation for cloud games: Are we there yet?" in *Proc. 13th Annu. Workshop Netw. Syst. Support Games*, Nagoya, Japan, Dec. 2014, pp. 1–6.

[37] M. S. Hossain, G. Muhammad, B. Song, M. Hassan, A. Alelaiwi, and A. Alamri, "Audio–visual emotion-aware cloud gaming framework," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 25, no. 12, pp. 2105–2118, Dec. 2015.

[38] C.-Y. Huang, K.-T. Chen, D.-Y. Chen, H.-J. Hsu, and C.-H. Hsu, "GamingAnywhere: The first open source cloud gaming system," *ACM Trans. Multimedia Comput., Commun., Appl.*, vol. 10, no. 1s, pp. 10:1–10:25, Jan. 2014.

[39] C.-Y. Huang, P.-H. Chen, Y.-L. Huang, K.-T. Chen, and C.-H. Hsu, "Measuring the client performance and energy consumption in mobile cloud gaming," in *Proc. 13th Annu. Workshop Netw. Syst. Support Games (NetGames)*, Nagoya, Japan, Dec. 2014, pp. 1–3.

[40] C.-Y. Huang, C.-H. Hsu, Y.-C. Chang, and K.-T. Chen, "GamingAnywhere: An open cloud gaming system," in *Proc. ACM Multimedia Syst. Conf. (MMSys)*, Oslo, Norway, Feb. 2013, pp. 36–47.

[41] C.-Y. Huang, C.-H. Hsu, D.-Y. Chen, and K.-T. Chen, "Quantifying user satisfaction in mobile cloud games," in *Proc. Workshop Mobile Video Del. (MoVid)*, Singapore, Mar. 2013, pp. 4:1–4:6.

[42] M. Jarschel, D. Schlosser, S. Scheuring, and T. Hoßfeld, "Gaming in the clouds: QoE and the users' perspective," *Math. Comput. Model.*, vol. 57, nos. 11–12, pp. 2883–2894, Jun. 2013.

[43] S. Jarvinen, J. P. Laulajainen, T. Sutinen, and S. Sallinen, "QoS-aware real-time video encoding: How to improve the user experience of a gaming-on-demand service," in *Proc. IEEE Consum. Commun. Netw. Conf. (CCNC)*, Las Vegas, NV, USA, Jan. 2006, pp. 994–997.

[44] K. I. Kim, S. Y. Bae, D. C. Lee, C. S. Cho, H. J. Lee, and K. C. Lee, "Cloud-based gaming service platform supporting multiple devices," *ETRI J.*, vol. 35, no. 6, pp. 960–968, Dec. 2013.

[45] S. S. Kim, K. I. Kim, and J. Won, "Multi-view rendering approach for cloud-based gaming services," in *Proc. Int. Conf. Adv. Future Internet (AFIN)*, French Riviera, France, Aug. 2011, pp. 102–107.

[46] U. Lampe, Q. Wu, S. Dargutev, R. Hans, A. Miede, and R. Steinmetz, "Assessing latency in cloud gaming," in *Proc. Int. Conf. Cloud Comput. Services Sci. (CLOSER)*, Barcelona, Spain, Sep. 2014, pp. 52–68.

[47] J. Laulajainen, T. Sutinen, and S. Järvinen, "Experiments with QoS-aware gaming-on-demand service," in *Proc. Int. Conf. Adv. Inf. Netw. Appl. (AINA)*, Vienna, Austria, Apr. 2006, pp. 805–810.

[48] K. Lee *et al.*, "Outatime: Using speculation to enable low-latency continuous interaction for cloud gaming," in *Proc. Annu. Int. Conf. Mobile Syst., Appl., Services (MobiSys)*, Florence, Italy, May 2015, pp. 151–165.

[49] K. Lee, D. Chu, E. Cuervo, A. Wolman, and J. Flinn, "Demo: DeLorean: Using speculation to enable low-latency continuous interaction for mobile cloud gaming," in *Proc. Annu. Int. Conf. Mobile Syst., Appl., Services (MobiSys)*, Florence, Italy, May 2015, p. 347.

[50] Y.-T. Lee, K.-T. Chen, H.-I. Su, and C.-L. Lei, "Are all games equally cloud-gaming-friendly? An electromyographic approach," in *Proc. 11th Annu. Workshop Netw. Syst. Support Games (NetGames)*, Venice, Italy, Nov. 2012, pp. 1–6.

[51] Y. Li, X. Tang, and W. Cai, "Play request dispatching for efficient virtual machine usage in cloud gaming," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 25, no. 12, pp. 2052–2063, Dec. 2015.

[52] L. Lin *et al.*, "LiveRender: A cloud gaming system based on compressed graphics streaming," in *Proc. ACM Int. Conf. Multimedia (MM)*, Orlando, FL, USA, Nov. 2014, pp. 347–356.

[53] Y. Liu, S. Dey, and Y. Lu, "Enhancing video encoding for cloud gaming using rendering information," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 25, no. 12, pp. 1960–1974, Dec. 2015.

[54] Y. Liu, S. Wang, and S. Dey, "Modeling, characterizing, and enhancing user experience in cloud mobile rendering," in *Proc. Int. Conf. Comput., Netw. Commun. (ICNC)*, Maui, HI, USA, Jan. 2012, pp. 739–745.

[55] M. Manzano, J. A. Hernandez, M. Uruena, and E. Calle, "An empirical study of cloud gaming," in *Proc. 11th IEEE Annu. Workshop Netw. Syst. Support Games (NetGames)*, Venice, Italy, Nov. 2012, pp. 1–2.

[56] M. Manzano, M. Urueña, M. Sužnjević, E. Calle, J. Hernández, and M. Matijasevic, "Dissecting the protocol and network traffic of the OnLive cloud gaming platform," *Multimedia Syst.*, vol. 20, no. 5, pp. 451–470, Mar. 2014.

[57] M. Marzolla, S. Ferretti, and G. D'Angelo, "Dynamic resource provisioning for cloud-based gaming infrastructures," *ACM Comput. Entertainment*, vol. 10, no. 3, pp. 4:1–4:20, Dec. 2012.

[58] D. Meilander, F. Glinka, S. Gorlatch, L. Lin, W. Zhang, and X. Liao, "Bringing mobile online games to clouds," in *Proc. IEEE Conf. Comput. Commun. Workshops (INFOCOMW)*, Toronto, ON, Canada, Apr. 2014, pp. 340–345.

[59] D. Mishra, M. El Zarki, A. Erbad, C.-H. Hsu, and N. Venkatasubramanian, "Clouds+Games: A multifaceted approach," *IEEE Internet Comput.*, vol. 18, no. 3, pp. 20–27, May 2014.

[60] S. Möller, D. Pommer, J. Beyer, and J. Rake-revelant, "Factors influencing gaming QoE: Lessons learned from the evaluation of cloud gaming services," in *Proc. Int. Workshop Perceptual Quality Syst. (PQS)*, Vienna, Austria, Sep. 2013, pp. 1–5.

[61] C. Moreno, N. Tizon, and M. Preda, "Mobile cloud convergence in GaaS: A business model proposition," in *Proc. Hawaii Int. Conf. Syst. Sci. (HICSS)*, Maui, HI, USA, Jan. 2012, pp. 1344–1352.

[62] (Mar. 2010). *Welcome to Native Client*. [Online]. Available: https://developer.chrome.com/native-client

[63] (Mar. 2010). *Google's Native Client Goes ARM and Beyond*. [Online]. Available: http://www.h-online.com/open/news/item/Google-s-Native-Client-goes-ARM-and-beyond-957478.html

[64] X. Nan *et al.*, "A novel cloud gaming framework using joint video and graphics streaming," in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, Chengdu, China, Jul. 2014, pp. 1–6.

[65] (May 2016). *NVidia Grid*. [Online]. Available: http://www.nvidia.com/object/cloud-gaming.html

[66] A. Ojala and P. Tyrvainen, "Developing cloud business models: A case study on cloud gaming," *IEEE Softw.*, vol. 28, no. 4, pp. 42–47, Jul. 2011.

[67] (Jan. 2015). *OnLive*. [Online]. Available: http://www.onlive.com/

[68] (Jan. 2015). *PlayStation Now*. [Online]. Available: http://www.playstation.com/en-us/explore/playstationnow/

[69] S. Prabu and S. Purushotham, "Cloud gaming with P2P network using XAML and Windows Azure," in *Proc. Conf. Recent Trends Comput., Commun. Inf. Technol. (ObCom)*, Vellore, India, Dec. 2011, pp. 165–172.

[70] Z. Qi, J. Yao, C. Zhang, M. Yu, Z. Yang, and H. Guan, "VGRIS: Virtualized GPU resource isolation and scheduling in cloud gaming," *ACM Trans. Archit. Code Optim.*, vol. 11, no. 2, pp. 203–214, Jul. 2014.

[71] P. Quax, A. Beznosyk, W. Vanmontfort, R. Marx, and W. Lamotte, "An evaluation of the impact of game genre on user experience in cloud gaming," in *Proc. IEEE Int. Games Innov. Conf. (IGIC)*, Vancouver, BC, Canada, Sep. 2013, pp. 216–221.

[72] K. Raaen, R. Eg, and C. Griwodz, "Can gamers detect cloud delay?" in *Proc. Annu. Workshop Netw. Syst. Support Games (NetGames)*, Nagoya, Japan, Dec. 2014, pp. 1–3.

[73] L. Riungu-kalliosaari, J. Kasurinen, and K. Smolander, "Cloud services and cloud gaming in game development," in *Proc. IADIS Game Entertainment Technol. (GET)*, Prague, Czech Republic, Jul. 2013, pp. 1–10.

[74] P. E. Ross, "Cloud computing's killer app: Gaming," *IEEE Spectr.*, vol. 46, no. 3, p. 14, Mar. 2009.

[75] (Nov. 2014). *Cloud Gaming to Reach Inflection Point in 2015*. [Online]. Available: http://tinyurl.com/p3z9hs2

[76] M. Semsarzadeh, M. Hemmati, A. Javadtalab, A. Yassine, and S. Shirmohammadi, "A video encoding speed-up architecture for cloud gaming," in *Proc. IEEE Int. Conf. Multimedia Expo Workshops*, Chengdu, China, Jul. 2014, pp. 1–6.

[77] M. Semsarzadeh, A. Yassine, and S. Shirmohammadi, "Video encoding acceleration in cloud gaming," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 25, no. 12, pp. 1975–1987, Dec. 2015.

[78] R. Shea, D. Fu, and J. Liu, "Cloud gaming: Understanding the support from advanced virtualization and hardware," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 25, no. 12, pp. 2026–2037, Dec. 2015.

[79] R. Shea, D. Fu, and J. Liu, "Towards bridging online game playing and live broadcasting: Design and optimization," in *Proc. ACM Workshop Netw. Oper. Syst. Support Digit. Audio Video (NOSSDAV)*, Portland, OR, USA, Mar. 2015, pp. 61–66.

[80] R. Shea and J. Liu, "On GPU pass-through performance for cloud gaming: Experiments and analysis," in *Proc. Annu. Workshop Netw. Syst. Support for Games (NetGames'13)*, p. 6:1–6:6, Denver, CO, Dec. 2013.

[81] R. Shea, J. Liu, E. C.-H. Ngai, and Y. Cui, "Cloud gaming: Architecture and performance," *IEEE Netw.*, vol. 27, no. 4, pp. 16–21, Jul./Aug. 2013.

[82] S. Shi, C.-H. Hsu, K. Nahrstedt, and R. Campbell, "Using graphics rendering contexts to enhance the real-time video coding for mobile cloud gaming," in *Proc. ACM Multimedia (MM)*, Nov. 2011, pp. 103–112.

[83] T. Kämäräinen, M. Siekkinen, Y. Xiao, and A. Ylä-Jääski, "Towards pervasive and mobile gaming with distributed cloud infrastructure," in *Proc. Annu. Workshop Netw. Syst. Support Games (NetGames)*, Nagoya, Japan, Dec. 2014, pp. 1–6.

[84] I. Slivar, M. Suznjevic, L. Skorin-Kapov, and M. Matijasevic, "Empirical QoE study of in-home streaming of online games," in *Proc. Annu. Workshop Netw. Syst. Support Games (NetGames)*, Nagoya, Japan, Dec. 2014, pp. 1–6.

[85] O. Soliman, A. Rezgui, H. Soliman, and N. Manea, "Mobile cloud gaming: Issues and challenges," in *Proc. Int. Conf. Mobile Web Inf. Syst. (MobiWIS)*, Paphos, Cyprus, Aug. 2013, pp. 121–128.

[86] (Jul. 2012). *Cloud Gaming Adoption is Accelerating … and Fast!* [Online]. Available: http://www.nttcom.tv/2012/07/09/cloud-gaming-adoption-is-acceleratingand-fast/

[87] (Apr. 2015). *Sony buys OnLive Streaming Game Service, Which Will Shut Down Later This Month*. [Online]. Available: http://goo.gl/6Xe9Dx

[88] K. Sun and D. Wu, "Video rate control strategies for cloud gaming," *J. Vis. Commun. Image Represent.*, vol. 30, pp. 234–241, Jul. 2015.

[89] M. Suznjevic, J. Beyer, L. Skorin-Kapov, S. Moller, and N. Sorsa, "Towards understanding the relationship between game type and network traffic for cloud gaming," in *Proc. IEEE Int. Conf. Multimedia Expo Workshops (ICMEW)*, Chengdu, China, Jul. 2014, pp. 1–6.

[90] R. Süselbeck, G. Schiele, and C. Becker, "Peer-to-peer support for low-latency massively multiplayer online games in the cloud," in *Proc. 8th Annu. Workshop Netw. Syst. Support Games (NetGames)*, Paris, France, Nov. 2009, pp. 1–2.

[91] M. R. Hosseinzadeh Taher, H. Ahmadi, and M. R. Hashemi, "Power-aware analysis of H.264/AVC encoding parameters for cloud gaming," in *Proc. IEEE Int. Conf. Multimedia Expo Workshops (ICMEW)*, Chengdu, China, Jul. 2014, pp. 1–6.

[92] H. Tian, D. Wu, J. He, Y. Xu, and M. Chen, "On achieving cost-effective adaptive cloud gaming in geo-distributed data centers," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 25, no. 12, pp. 2064–2077, Dec. 2015.

[93] (Jan. 2015). *Ubitus*. [Online]. Available: http://www.ubitus.net

[94] S. Wang and S. Dey, "Modeling and characterizing user experience in a cloud server based mobile gaming approach," in *Proc. IEEE Global Telecommun. Conf. (GLOBECOM)*, Honolulu, HI, USA, Nov. 2009, pp. 1–7.

[95] S. Wang and S. Dey, "Addressing response time and video quality in remote server based Internet mobile gaming," in *Proc. IEEE Wireless Commun. Netw. Conf.*, Sydney, NSW, Australia, Apr. 2010, pp. 1–6.

[96] S. Wang and S. Dey, "Rendering adaptation to address communication and computation constraints in cloud mobile gaming," in *Proc. IEEE Global Telecommun. Conf. (GLOBECOM)*, Miami, FL, USA, Dec. 2010, pp. 1–6.

[97] S. Wang and S. Dey, "Cloud mobile gaming: Modeling and measuring user experience in mobile wireless networks," *ACM Trans. Mobile Comput. Commun. Rev.*, vol. 16, no. 1, pp. 10–21, Jan. 2012.

[98] S. Wang, Y. Liu, and S. Dey, "Wireless network aware cloud scheduler for scalable cloud mobile gaming," in *Proc. IEEE Int. Conf. Commun. (ICC)*, Ottawa, ON, Canada, Jun. 2012, pp. 2081–2086.

[99] J. Wu, C. Yuen, N.-M. Cheung, J. Chen, and C. W. Chen, "Enabling adaptive high-frame-rate video streaming in mobile cloud gaming applications," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 25, no. 12, pp. 1988–2001, Dec. 2015.

[100] Z. Wu, "Gaming in the cloud: One of the future entertainment," in *Proc. Interact. Multimedia Conf.*, Southampton, U.K., Jan. 2014, pp. 1–6.

[101] L. Xu, X. Guo, Y. Lu, S. Li, O. Au, and L. Fang, "A low latency cloud gaming system using edge preserved image homography," in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, Chengdu, China, Jul. 2014, pp. 1–6.

[102] Z. Xue, D. Wu, J. He, X. Hei, and Y. Liu, "Playing high-end video games in the cloud: A measurement study," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 25, no. 12, pp. 2013–2025, Dec. 2015.

[103] C. Zhang, Z. Qi, J. Yao, M. Yu, and H. Guan, "vGASA: Adaptive scheduling algorithm of virtualized GPU resource in cloud gaming," *IEEE Trans. Parallel Distrib. Syst.*, vol. 25, no. 11, pp. 3036–3045, Nov. 2014.

[104] Z. Zhao, K. Hwang, and J. Villeta, "Game cloud design with virtualized CPU/GPU servers and initial performance results," in *Proc. Workshop Sci. Cloud Comput. Date (ScienceCloud)*, Delft, The Netherlands, Jun. 2012, pp. 23–30.

WEI CAI [S'12–M'16] received the B.Eng. degree in software engineering from Xiamen University in 2008, the M.S. degree in electrical engineering and computer science from Seoul National University in 2011, and the Ph.D. degree in electrical and computer engineering from The University of British Columbia (UBC), Vancouver, BC, Canada, in 2016. He is currently a Post-Doctoral Research Fellow with the UBC. He has completed visiting researches with Academia Sinica, The Hong Kong Polytechnic University, and National Institute of Informatics, Japan. His researches focus on gaming as a service, mobile cloud computing, online gaming, software engineering, and interactive multimedia. He received awards of the 2015 Chinese Government Award for the Outstanding Self-Financed Students Abroad, UBC Doctoral Four-Year-Fellowship, Brain Korea 21 Scholarship, and Excellent Student Scholarship from the Bank of China. He is also a co-recipient of the best paper awards from the CloudCom2014, the SmartComp2014, and the CloudComp2013.

RYAN SHEA [S'08–M'16] received the Ph.D. degree in computer science from Simon Fraser University, Burnaby, BC, Canada, in 2016. He is currently a University Research Associate with the Big Data Systems. His research interests include computer and network virtualization, performance issues in cloud computing, and energy and performance issues with the Big Data Systems. He has received the Natural Sciences and Engineering Research Council of Canada, Alexander Graham Bell Canada Graduate Scholarship in 2013. He was a recipient of the best student paper award from the IEEE/ACM 21st International Workshop on Quality of Service for his paper Understanding the Impact of Denial of Service Attacks on Virtual Machines in 2012. His recent publications include a point of view article in the Proceedings of the IEEE entitled The Future of Cloud Gaming.

CHUN-YING HUANG [S'03–M'08] received the Ph.D. degree in electrical engineering department from National Taiwan University in 2007. He leads the Security and Systems Laboratory with National Chiao Tung University. From 2008 to 2013, he was an Assistant Professor with the Department of Computer Science and Engineering, National Taiwan Ocean University, where he was an Associate Professor from 2013 to 2016. He has been an Associate Professor with the Department of Computer Science, National Chiao Tung University, since 2016. His research interests include system security, multimedia networking, and mobile computing. He is a member of ACM.

KUAN-TA CHEN [S'04–M'06–SM'15] received the B.S. and M.S. degrees in computer science from National Tsing-Hua University, in 1998 and 2000, respectively, and the Ph.D. degree in electrical engineering from National Taiwan University in 2006. He is a Research Fellow with the Institute of Information Science, and the Research Center for Information Technology Innovation (joint appointment) of Academia Sinica. His research interests include quality of experience, multimedia systems, and social computing. He received the best paper award in the IWSEC 2008 and the K. T. Li Distinguished Young Scholar Award from ACM Taipei/Taiwan Chapter in 2009. He also received the Outstanding Young Electrical Engineer Award from The Chinese Institute of Electrical Engineering in 2010, the Young Scholar's Creativity Award from Foundation for the Advancement of Outstanding Scholarship in 2013, and the IEEE ComSoc MMTC Best Journal Paper Award in 2014. He was an Associate Editor of IEEE Transactions on Multimedia from 2011 to 2014 and has been an Associate Editor of *ACM Transactions on Multimedia Computing, Communications, and Applications* since 2015. He is a Senior Member of ACM.

JIANGCHUAN LIU [S'01–M'03–SM'08] received the B.Eng. (*cum laude*) degree in computer science from Tsinghua University, Beijing, China, in 1999, and the Ph.D. degree in computer science from The Hong Kong University of Science and Technology in 2003.

He was an Assistant Professor with The Chinese University of Hong Kong from 2003 to 2004. He is a University Professor with the School of Computing Science, Simon Fraser University, British Columbia, Canada, and an NSERC E.W.R. Steacie Memorial Fellow. He is an EMC-Endowed Visiting Chair Professor of Tsinghua University, Beijing, from 2013 to 2016. His research interests include multimedia systems and networks, cloud computing, social networking, online gaming, big data computing, wireless sensor networks, and peer-to-peer networks.

Dr. Liu is a co-recipient of the inaugural Test of Time Paper Award of the IEEE INFOCOM (2015), the ACM SIGMM TOMCCAP Nicolas D. Georganas best paper award (2013), the ACM Multimedia best paper award (2012), the IEEE Globecom best paper award (2011), and the IEEE Communications Society best paper award on Multimedia Communications (2009). His students received the Best Student Paper Award of the IEEE/ACM IWQoS in 2008 and 2012. He has served on the editorial boards of the IEEE Transactions on Big Data, the IEEE Transactions on Multimedia, the IEEE Communications Surveys and Tutorials, the IEEE Access, the IEEE Internet of Things Journal, *Computer Communications*, and *Wiley Wireless Communications and Mobile Computing*. He is the Steering Committee Chair of the IEEE/ACM IWQoS (2015–2017) and TPC Co-Chair of the IEEE IC2E'2017, and the IEEE/ACM IWQoS'2014. He serves as an Area Chair of the IEEE INFOCOM, ACM Multimedia, and the IEEE ICME. According to Google Scholar, the citations of his papers are over 10,000, and as an h-index of 46.

VICTOR C. M. LEUNG [S'75–M'89–SM'97–F'03] is a Professor of Electrical and Computer Engineering and holder of the TELUS Mobility Research Chair with The University of British Columbia (UBC). His research is in the areas of wireless networks and mobile systems. He has co-authored over 900 technical papers in archival journals and refereed conference proceedings, several of which had received best paper awards. He is a fellow of the Royal Society of Canada, the Canadian Academy of Engineering, and the Engineering Institute of Canada. He is serving or has served on the editorial boards of the IEEE Access, the IEEE Journal on Selected Areas in Communications – Series on Green Communications and Networking, the IEEE Wireless Communications Letters, the IEEE Transactions on Computers, the IEEE Transactions on Wireless Communications, and the IEEE Transactions on Vehicular Technology, and several other journals. He has provided leadership to the technical program committees and organizing committees of numerous international conferences. He was the recipient of the 1977 APEBC Gold Medal, NSERC Postgraduate Scholarships from 1977 to 1981, 2012 UBC Killam Research Prize, and the IEEE Vancouver Section Centennial Award.

CHENG-HSIN HSU [S'09–M'10–SM'16] received M.S./B.S. degrees from National Chung-Cheng University, the M.Eng. degree from the University of Maryland, and the Ph.D. degree from Simon Fraser University. He was with the Deutsche Telekom Laboratory, Motorola Inc., and Lucent Technologies. He has been an Associate Professor with the Department of Computer Science, National Tsing Hua University, since 2014, where he was an Assistant Professor from 2011 to 2014. He visited the University of California at Irvine, Qatar Computing Research Institute, and University of Illinois Urbana–Champaign, in 2013, 2014, and 2015, respectively. His research interests are in multimedia networking, mobile computing, and computer networks. He has been served as an Associate Editor of *ACM Transactions on Multimedia Computing, Communications, and Applications* since 2014, and the IEEE MMTC E-Letter from 2012 to 2014. He and his colleagues received the best paper award from the IEEE RTAS'12, the TAOS best paper award from the IEEE GLOBECOM'12, best paper award from the IEEE Innovation'08, and the Best Demo Award from the ACM Multimedia'08.