Ulrich Lampe, Qiong Wu, Sheip Dargutev, Ronny Hans, André Miede, Ralf Steinmetz: Assessing Latency in Cloud Gaming. In: Markus Helfert, Frédéric Desprez, Donald Ferguson, Frank Leymann: Cloud Computing and Services Science, vol. CCNS, 453, chap. 4, p. 52-68, Springer International Publishing, September 2014. ISBN 978-3-319-11560-3.

Assessing Latency in Cloud Gaming

Ulrich Lampe[1], Qiong Wu[1], Sheip Dargutev[1], Ronny Hans[1], André Miede[2], and Ralf Steinmetz[1]

[1] Multimedia Communications Lab (KOM), TU Darmstadt, Rundeturmstr. 10, 64283 Darmstadt, Germany, {firstName.lastName}@KOM.tu-darmstadt.de
[2] Fakultät für Ingenieurwissenschaften, htw saar, Goebenstr. 40, 66117 Saarbrücken, Germany, andre.miede@htwsaar.de

Abstract. With the emergence of cloud computing, diverse types of Information Technology services are increasingly provisioned through large data centers via the Internet. A relatively novel service category is cloud gaming, where video games are executed in the cloud and delivered to a client as audio/video stream. While cloud gaming substantially reduces the demand of computational power on the client side, thus enabling the use of thin clients, it may also affect the Quality of Service through the introduction of network latencies. In this work, we quantitatively examined this effect, using a self-developed measurement tool and a set of actual cloud gaming providers. For the two providers and three games in our experiment, we found absolute increases in latency between approximately 40 ms and 150 ms, or between 85% and 800% in relative terms, compared to a local game execution. In addition, based on a second complementary experiment, we found mean round-trip times ranging from about 30 ms to 380 ms using WLAN and approximately 40 ms to 1050 ms using UMTS between a local computer and globally distributed compute nodes. Bilaterally among the compute nodes, results were in the range from approximately 10 ms to 530 ms. This highlights the importance of data center placement for the provision of cloud gaming services with adequate Quality of Service properties.

1 Introduction

Since its popularization in the mid-2000s, cloud computing has substantially altered the way in which Information Technology (IT) services are delivered and brought massive changes to the IT sector [1]. Today, the decade-old vision of delivering IT as a "utility" has come closer to realization than ever before [2]. A relatively novel business model, within the greater context of cloud computing, is *cloud gaming*. The principal idea of this concept is to execute video games in a cloud data center and deliver them to a client as audio/video stream via the Internet. The client thus serves as a simple playback and input device; the computationally complex task of executing the actual game logic and rendering the game images is shifted to the cloud [3–6]. From a formal standpoint, based on the popular NIST definition of cloud computing [7], cloud gaming can most

intuitively be interpreted as a subclass of the *Software as a Service* model, because it constitutes a functionally complex service that is offered on the basis of low-level infrastructure services.

From a customer perspective, one main advantage of cloud gaming exists in the ability to access games at any place and time, independent of any specific device upon which they are installed [3]. Furthermore, hardware expenditures are substantially reduced, because a simplistic thin client is usually sufficient for access [8]. In addition, games do not have to be purchased for a fixed (and commonly quite notable) amount of money, but can be leased on a pay-per-use basis. From the provider perspective, one main benefit is the prevention of copyright infringements [5]. Furthermore, the development process may be greatly simplified if games are exclusively developed for the cloud, rather than multiple different platforms.

However, the use of the Internet also introduces a new component into the delivery chain. Being a public network, the Internet lies (partially) out of the control sphere of both the user and the provider, and follows a "best effort" philosophy, i. e., it does not make any end-to-end Quality of Service (QoS) assurances [9]. Hence, limitations of the network infrastructure, such as high latency, small bandwidth, or high packet loss, may potentially affect the QoS of the *overall* cloud gaming system for the user.

In this work, we focus on the QoS parameter of latency. This parameter plays an important role for the overall game experience [10, 6]. This specifically applies for action-oriented games such as first-person shooters, where it may determine whether a player is "fragged", i.e., her/his character is killed, or is able to frag her/his opponent [11, 10]. Hence, the first research question we aim to empirically answer in this work is: "What is the impact of cloud gaming on the QoS parameter of latency, as compared to a local execution of a video game?". In addition, inspired by related research [3] and our own previous work [12], we formulate the following second research question: "What is the impact of the geographical placement of cloud data centers on the QoS parameter of latency?".

In the following Section 2, we address the first research question, concerning the difference between cloud-based and local gaming. This includes an extensive description of our procedure and a thorough presentation and discussion of results. The second research question, concerning the latency implications of global data center placement, is addressed in Section 3. An overview of related work is given in Section 4. The paper closes with a summary and general conclusions in Section 5.

2 Examination of Latency in Cloud-Based and Local Gaming

In this section, we describe the first part of our experiments, aiming at the quantification of latencies in cloud-based and local gaming. Following a description of

the considered variables, the measurement tool, and the measurement procedure, we present our results and along with a thorough discussion.

2.1 Considered Variables

As explained in the previous section, in this work, we focus on the QoS parameter of latency. It thus constitutes the only *dependent* variable in our experiments. More specifically, we consider *user-perceived latency*. By that term, we refer to the timespan that elapses between a certain *action* performed by the user, e. g., the press of a mouse button or a key, and the corresponding game *reaction*, e. g., the appearance of gunfire or the menu. It is also referred to as "interactive response time" or "response delay" in related research [3, 13]. Based on the combined findings of Choy et al., Wang, and Wilson latency can be split into the following components if a game is locally executed [3, 14, 15]:

- *Input lag*, which corresponds to the timespan between two subsequent sampling events of the game controller, e. g., mouse or keyboard.
- *Game pipeline CPU time*, i. e., the time which is required for processing the input and realizing the game logic.
- *Game pipeline GPU time*, i. e., the time which the graphic card requires for rendering the next frame of the game.
- *Frame transmission*, which denotes the time that is required for transferring the frame from the backbuffer to the frontbuffer of the graphic card, and subsequently to the screen.
- *LCD response time*, which indicates the timespan that is required to actually display the frame on the screen.

Once a game is executed in the cloud and delivered via a network, the following additional components have to be considered [3, 14, 15]:

- *Upstream data transfer*, i. e., the time that it takes to sent the user input to the cloud gaming provider.
- *Capture and encoding*, which denotes the time requirements for capturing the current frame and encoding it as video stream.
- *Downstream data transfer*, i. e., the timespan for transferring the stream to the client.
- *Decoding*, which indicates the time for converting the video stream back into a frame.

Intuitively, one might reason that a cloud-based game will always exhibit a higher latency that a locally executed game due to the additional latency components. However, this is not necessarily true. In fact, due to the use of potent hardware in the cloud and depending on the geographical distance between the user and the cloud provider, the reduction of time spent in the game pipeline may overcompensate the network, encoding, and decoding latencies [14].

The dependent variable in our experiment, latency, may potentially be determined by various factors, i. e., a set of *independent* variables. In our work, we

focus on different games, cloud gaming providers, and network connections as suspected key determinants.

With respect to the main subject of our research, i. e., the examined games, our focus was on action-oriented titles. As briefly explained in Section 1, these games are commonly very sensitive to latency increases and thus, of elevated interest. We specifically chose the following titles, all of which are available both in the cloud and for local installation:

- *Shadowgrounds*[3] is a 3D first-person shooter game developed by Frozenbyte. It was initially released in the year 2005.
- *Shadowgrounds Survivor*[4] is a sequel to Shadowgrounds. It was also developed by Frozenbyte and released in 2007.
- *Trine*[5] is an action-oriented puzzle game. It was developed by Frozenbyte as well and released in 2009.

The determination of representative cloud gaming providers is somewhat challenging. Following an initial hype around cloud gaming, which resulted in a variety of new suppliers, the market appears to be in a phase of consolidation today. For example, *Gaikai*, one of the pioneers in cloud gaming, was acquired in August 2012 by the major industry player *Sony* [16], and had temporally ceased its services; recent reports indicate that Sony plans to exploit Gaikai for the delivery of games to its new *PlayStation 4* gaming console starting in the third quarter of 2014 [17]. This work includes measurements for three provisioning options:

- *Cloud Gaming Provider A* (CGP-A), which is located in the Americas and operates a dedicated infrastructure[6].
- *Cloud Gaming Provider B* (CGP-B), with headquarters in the Asian-Pacific region, which also uses a dedicated infrastructure.
- *A local personal computer* (Local), which is equipped with an Intel Core 2 Quad Q6700 CPU, an NVidia Geforce GTX 560 GPU, and 4 GB of memory.

As it has been explained before, cloud gaming employs the Internet as delivery channel. Because the network as such is out of the control sphere of both provider and user, we focus on the user's network connection in our experiments. Specifically, we regard the following techniques:

- *Universal Mobile Telecommunications System* (UMTS), which marks the third generation (3G) of cellular networks and has been widely deployed in many industrialized countries since the mid-2000s. We use a variant with the *High Speed Packet Access* (HSPA) extensions.

[3] `http://www.shadowgroundsgame.com/`
[4] `http://www.shadowgroundssurvivor.com/`
[5] `http://www.trine-thegame.com/`
[6] Unfortunately, due to legal considerations, we are required to anonymize the names of the cloud gaming providers.

– *Long Term Evolution* (LTE), which corresponds to the fourth generation (4G) of cellular networks. It has recently been or is currently being introduced by many mobile network providers.
– *Very High Speed Digital Subscriber Line* (VDSL), which denotes the cutting-edge in traditional fixed-line, copper cable-based Internet access.

2.2 Measurement Tool and Procedure

The aim of our approach is to automate the measurement process to the largest possible extent. For that matter, we have devised a GAme LAtency MEasurement TOol, or in brief, GALAMETO.KOM. This tool autonomously invokes a predefined action in the game and measures the time interval until the corresponding reaction can be observed. As a preparatory step, the tool requires the user to specify the trigger that invokes a certain action in the game. Such trigger may consist in pressing a mouse button or a key. Furthermore, the user has to specify the screen area that will reflect the corresponding reaction, such as the display of gunfire or the main menu. In order to reliably identify the reaction, the user further declares a numerical sensitivity value δ. This sensitivity value reflects the change of the *average color* within the predefined screen area. Lastly, in order to start an experiment, the user specifies the desired number of observations in the sample.

For each measurement iteration, GALAMETO.KOM first invokes the specified trigger. That is, it submits the user-defined activity to the game and stores a timestamp t_{act}. Then, the tool scans the frontbuffer of the graphics card and computes the initial average color value c_{init} for the predefined screen area. That procedure is continuously repeated, each time updating the current average color c_{curr} and a corresponding timestamp t_{react}. Once a change of color, i. e., a reaction with sufficient magnitude, is detected (i. e., if $|c_{curr} - c_{init}| \geq \delta$ holds), the latency $t_{lat} = t_{react} - t_{act}$ can be computed. The latency value is stored as new observation, and the process is repeated until a sample of the desired size has been collected.

For our experiment, we followed a so-called *full factorial design*. That is, we conducted measurements for each possible value combination of the three independent variables. Because the local execution of a single-player game is independent of the network connection, there are seven possible combinations of provider and network. For each combination, we examine the three selected games. Thus, our experimental setup consists of 21 different *test cases*.

For each test case, we acquired a sample of 250 observations. Subsequently, we checked for statistically significant differences between the test cases with respect to the mean latencies using a parametric t-test [18, 19]. For validation purposes, a non-parametric Mann-Whitney U-test was additionally applied [19]. Both tests were conducted at the same confidence level of 95% (i. e., $\alpha = 0.05$). The mean latencies of a pair of test cases are only considered significantly different if the according indication is given by both tests.

All experiments were executed using the previously specified computer in order to avoid measurement inaccuracies due to hardware differences. The differ-

ent network connections were provided by a major German telecommunications provider. No artificial network disturbances were introduced into the measurement process.

2.3 Results and Discussion

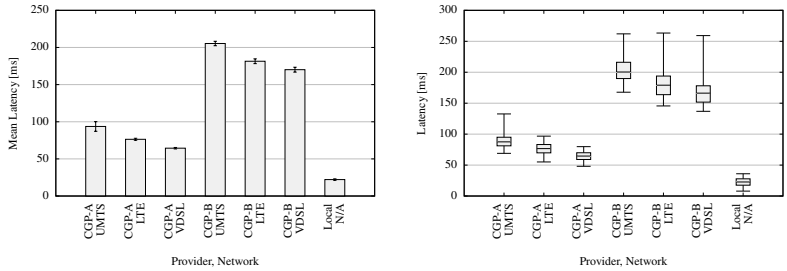
The results of our experiment, i. e., observed mean latencies are illustrated in Figures 1, 2, and 3 for the three games respectively. In addition, Table 1 contains the detailed results that have been the basis for the figures.

As can be seen, a local execution of the games yields the lowest latencies, ranging from 22 ms for Shadowgrounds to 44 ms for Trine. As it may have been expected, the latencies significantly increase with the novelty of the game. Because the remaining latency components can be assumed constant, this indicates a growth of computational complexity within the game pipeline, i. e., the overall increase in latency can likely be traced back to increased CPU and GPU time.

For cloud gaming provider A, we observe mean latencies between approximately 65 ms and 130 ms. The latencies significantly decrease with improved network connectivity. Specifically, with respect to the cellular networks, LTE is able to reduce the mean latency by up to 35 ms compared to UMTS. A fixed-line connection, namely VDSL, yields a further reduction of up to 12 ms. In general, the latency increases diminish compared to a local execution with the novelty of the game. This indicates that the latency of the game pipeline can, in fact, be reduced through the use of dedicated hardware in the cloud data center (cf. Section 2.1). However, the effect does not compensate for the network delay in our test cases. Hence, regardless of the game and network connection, provider A is not able to compete with a local execution in terms of latency. Depending on the network connection, cloud gaming adds between 40 ms and 90 ms of latency for each considered game. These differences are statistically significant at the assumed confidence level of 95%.

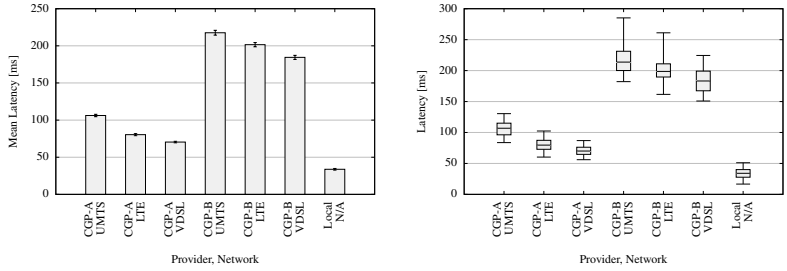
For cloud gaming provider B, we find even higher mean latencies between about 150 ms and 220 ms. Once again, there is a significant reduction in these figures with improved network connectivity. Compared to UMTS, LTE achieves a reduction of up to 29 ms, which very similar to the results for cloud gaming provider A. Likewise, VDSL shaves off between 9 ms and 17 ms in latency in comparison to LTE. In contrast to provider A, we do not find a decreasing latency margin with increasing novelty, i. e., computational complexity, of the game. Thus, provider B is even less capable than provider A of competing with a local execution in terms of latency. Specifically, depending on the game, provider B adds between 100 ms and 150 ms of latency. As for provider A, these increases are statistically significant.

In addition, the box-and-whisker plots indicate higher variations in the observed latencies, i. e., higher jitter, for both cloud gaming providers compared to a local game execution. This is also important to note, since previous research has not only identified absolute latencies, but also high jitter as an aspect in cloud gaming systems that may substantially impair the QoE of the end user [20].



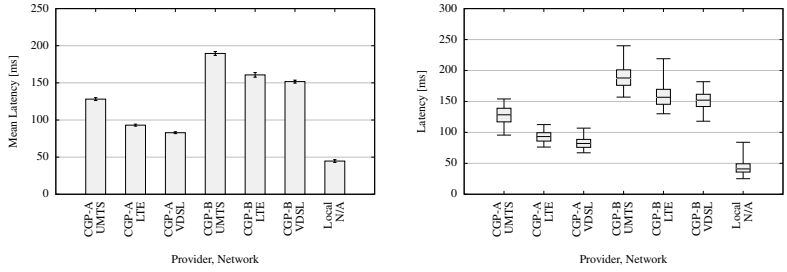
(a) Mean latencies. (b) Box-and-whisker plot of latencies.

Fig. 1. Observed latencies for the game *Shadowgrounds* (sample size $n = 250$). In Figure 1(a), error bars indicate the 95% confidence intervals. In Figure 1(b), the box marks the 25th, 50th, and 75th percentiles, and the whiskers indicate the 2.5th and 97.5th percentiles.



(a) Mean latencies. (b) Box-and-whisker plot of latencies.

Fig. 2. Observed latencies for the game *Shadowgrounds Survivor* (sample size $n = 250$). Same notation as in Figure 1.



(a) Mean latencies. (b) Box-and-whisker plot of latencies.

Fig. 3. Observed latencies for the game *Trine* (sample size $n = 250$). Same notation as in Figure 1.

Table 1. Detailed results for the independent variable latency per test case (in ms). Abbreviations: SG – Shadowgrounds; SGS – Shadowgrounds Survivor; CI95 – Radius of the 95% confidence interval; Pc. – Percentile.

Game	Provider	Network	Mean	CI95	2.5th Pc.	25th Pc.	Median	75th Pc.	97.5th Pc.
SG	CGP-A	UMTS	93.65	6.49	68.98	81.06	87.49	95.00	132.66
SG	CGP-A	LTE	76.39	1.34	55.04	69.74	76.65	83.19	96.76
SG	CGP-A	VDSL	64.39	0.98	48.01	59.05	64.52	69.96	79.75
SG	CGP-B	UMTS	205.34	3.00	167.71	189.81	200.33	216.11	262.00
SG	CGP-B	LTE	181.47	3.20	145.54	163.79	179.08	193.80	263.31
SG	CGP-B	VDSL	170.09	3.29	136.85	151.75	166.23	178.13	259.15
SG	Local	N/A	22.13	0.93	7.91	17.46	22.68	27.80	36.00
SGS	CGP-A	UMTS	106.19	1.61	83.63	96.21	106.89	115.02	130.41
SGS	CGP-A	LTE	80.41	1.40	60.26	72.82	79.59	87.32	102.33
SGS	CGP-A	VDSL	70.43	1.00	56.06	64.66	70.00	76.13	86.90
SGS	CGP-B	UMTS	217.63	3.27	182.11	200.11	213.73	231.18	285.12
SGS	CGP-B	LTE	201.58	2.85	161.56	189.64	198.71	210.90	261.06
SGS	CGP-B	VDSL	184.45	2.73	150.83	167.37	183.18	199.11	224.45
SGS	Local	N/A	33.79	1.11	16.64	27.69	34.03	39.98	51.08
Trine	CGP-A	UMTS	128.13	1.91	95.56	117.01	128.43	139.00	153.98
Trine	CGP-A	LTE	93.06	1.31	76.26	85.96	93.24	99.48	112.61
Trine	CGP-A	VDSL	82.88	1.25	67.05	75.82	82.03	88.62	106.85
Trine	CGP-B	UMTS	189.58	2.57	157.01	176.04	187.87	201.02	239.97
Trine	CGP-B	LTE	160.76	3.11	130.12	145.36	156.74	169.60	219.02
Trine	CGP-B	VDSL	151.69	2.01	118.01	141.79	152.10	161.56	181.86
Trine	Local	N/A	44.68	1.83	25.14	35.90	41.01	49.29	84.01

In summary, with respect to the first research question from Section 1, we conclude that cloud gaming has a significant and negative impact on the QoS parameter of latency, compared to the local execution of a game. Depending on the provider and network connection, cloud gaming results in latency increases between 40 ms and 150 ms. In relative terms, the increases amount to between 85% (Trine at CGP-A using VDSL) and 828% (Shadowgrounds at CGP-B using UMTS).

As previously explained, our focus in this work was on QoS, i. e., objective quality figures. Thus, the subjective perception of our results may substantially differ between various player groups. According to Dick et al., the mean tolerable latencies for an unimpaired experience in a multi-player game are in the range between 50 and 100 ms; maximal tolerable latencies are approximately 50 ms higher, i. e., in the order of 100 to 150 ms [10]. User studies by Jarschel et al. also indicate that the Quality of Experience (QoE) quickly drops with increasing latency, specifically in fast-paced games such as racing simulations or first-person shooters [4]. Hence, based on the observed numbers, we believe that cloud gaming is primarily attractive for slow-paced games, as well as casual players who likely have moderate QoS expectations compared to experienced and sophisticated gamers.

Given the reliance on the Internet as delivery medium, cloud gaming would likely profit from a shift away from the best-effort philosophy towards sophisticated QoS mechanisms. The development of such mechanisms has been an active field of research for many years, resulting in proposals such as *Integrated Services* (IntServ) or *Differentiated Services* (DiffServ) [21]. However, past experience – for example, with the rather sluggish introduction of *IPv6* – has shown that many Internet service providers are rather reluctant to make fundamental infrastructure changes unless a pressing need arises. In addition, as the ongoing

debate about *net neutrality* shows, the introduction of QoS management techniques on the Internet is not merely a technical issue. For a more comprehensive discussion, we refer the interested reader to Xiao [22].

Assuming that the Internet itself will remain to follow a best-effort philosophy in the short and medium term, two main options remain for cloud providers to improve the QoS of their systems. The first option consists in moving the data centers geographically closer to the clients. However, for a constant client base, such decentralization implies building a larger number of data centers. Due to the reduced size and thus, smaller economies of scale of these data centers [23], such approach is likely to be cost-intensive. A viable alternative may consist in the exploitation of servers in existing content delivery networks, as proposed by Choy et al. [3]. Second, cloud providers may upgrade their servers to reduce the latency of the game pipeline. Thus, they could aim to (over-)compensate for the network latency. However, while such an approach may be successful for computationally complex games, it will likely fail for older games where the impact of the game pipeline is relatively small. In addition, server upgrades can be costly, specifically if disproportionately expensive high-end components have to be purchased. Hence, in our opinion, a key challenge for cloud providers consists in finding an economically reasonable balance between QoS (and thus, the potential number of customers) and cost.

3 Examination of Round-Trip Times Between Globally Distributed Compute Nodes

This section describes the second part of our experiments, i. e., the quantification of latencies between globally distributed compute nodes, representing fictitious cloud data centers. Similar to the previous section, we first explain the considered variables and measurement procedure, followed by a presentation and discussion of results.

3.1 Considered Variables

For this second part of our research, we conceived two linked experiments. The first examines the *bilateral* latency among different globally distributed locations in order to assess the maximum feasible distance between a provider and consumer. In contrast, the second experiment focuses on the *unilateral* latency between a single client and aforementioned compute nodes, depending on different network connections.

Similar to the previous section, latency constitutes the dependent variable in our experiments. However, in contrast to our previous experiment, we focus on a specific component of latency, namely *Round-Trip Time* (RTT), i. e., the timespan between the sending of a ping packet and receipt of the corresponding pong packet. RTT is also referred to as "network delay" in related research [13]. Since RTTs are only a part of overall latency in cloud gaming systems –

cf. Section 2.1 – and processing times are explicitly not considered, our results should be seen as a lower bound on overall latency.

For both the bilateral and unilateral measurements, we used an identical set of 15 globally distributed compute nodes, each representing a fictitious cloud data center. For the bilateral measurements, the *source node* and *target node* constitute independent variables. In the unilateral measurements, only the latter is considered, because all measurements are taken from the identical source host. In the unilateral measurements, we further considered the client's *network connection* as independent variable. Since our experiments were conducted independent of the latency measurements from Section 2, we did not have access to the dedicated network connections of aforementioned telecommunications provider anymore. Hence, we used the following two options:

- *UMTS* with HSPA extensions, using the public cellular network of another major German telecommunications provider.
- *Wireless Local Area Network* (WLAN), provided by an access point that was attached to our university's high-bandwidth network.

3.2 Measurement Tool and Procedure

For the purposes of our experiment, we conceived and implemented an additional measurement utility. It uses the tool *tcping*[7] for Windows respectively a comparable script *tcpping*[8] for Linux to repeatedly measure the RTT between the respective node and all remaining nodes.

In accordance with the previous section, we also followed a *full factorial design*, conducting measurements for each distinct value combination of the independent variables. In the case of the bilateral measurements, this results in 210 test cases, each representing a specific combination of the 15 source and target nodes. For the unilateral measurements, we considered 30 test cases, each resulting from a distinct combination of the 15 target nodes and two network connections.

As common testbed for our experiments, we chose PlanetLab [24], a research network that consisted of approximately 1,200 globally distributed nodes at 550 sites as of January 2014[9]. We selected 15 nodes on five continents (Asia, Africa, America, Europe, Australia) to serve as fictitious cloud data centers, representing the providers of cloud-based multimedia services, and deployed our measurement utility on these nodes. The measurement utility was executed over a timespan of approximately four days, resulting in samples of 10,000 RTT observations for each pair of nodes in the bilateral measurements. For the unilateral measurements, we obtained a smaller sample of 1,000 RTT observations for each network connection and target node. For each sample, we subsequently computed the mean RTT value and corresponding 95% confidence interval [19].

[7] `http://www.elifulkerson.com/projects/tcping.php`
[8] `http://www.vdberg.org/~richard/tcpping.html`
[9] `http://www.planet-lab.org/`

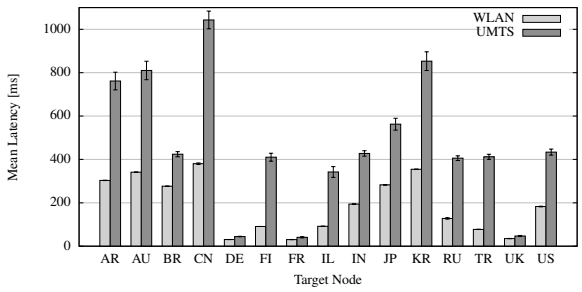


Fig. 4. Observed mean RTTs in the unilateral measurements (with 95% confidence intervals), using clients with different connection types in Darmstadt, Germany, by target node (effective sample size $n' \geq 991$).

3.3 Results and Discussion

The results for the bilateral and unilateral measurements are provided in Table 2 and Table 3. Furthermore, the results for the unilateral measurements are visualized in Figure 4. Please note that timed-out pings were not considered in the computation of mean RTTs, hence resulting in differing *effective* sizes for each underlying sample.

Since the propagation of electric signals is constrained by the speed of light, it comes as little surprise that the observed latencies dramatically increase with the geographical distance between two nodes. Choy et al. [3] explain that the processing of game content on the client and server side requires approximately 20 ms. This aligns well with the overall latency of approximately 20 ms to 40 ms that we found for local gaming in the previous section and also corresponds to delays of about 30 ms that were found by Huang et al. as part of experimental studies with their GamingAnywhere system [13]. Hence, RTTs should not exceed a threshold of about 80 ms to permit for acceptable QoE.

This argument confirms our statement from Section 2.3 that cloud data centers for the provision of cloud gaming services should be placed in geographical proximity to the potential clients. Specifically, the results for the bilateral measurements from Table 2 indicate the such proximity more or less translates into placement on the same continent. For example, RTTs between Argentina (AR) and Brazil (BR), or Germany (DE) and Finland (FI) were found to fairly accurately meet the aforementioned threshold of 80 ms.

This observation is underpinned by the results of the unilateral measurements in Table 3. Specifically in UMTS networks, the observed RTT drastically increases with geographical distance and easily exceeds the acceptable threshold, even for target nodes that were placed close to our client in Darmstadt, Germany, e. g., in Finland (FI) or Israel (IL). Furthermore, the results indicate that the gap in RTT between UMTS and WLAN is statistically significant at the assumed confidence level of 95% and quickly widens with geographical distance.

Table 2. Observed mean RTTs (in ms; with 95% confidence intervals) between pairs of globally distributed nodes (effective sample size $n' \geq 9,490$).

		Target Node													
		AR	AU	BR	CN	DE	FI	FR	IL	IN	JP	KR	RU	TR	UK	US
Source Node	AR	-	360.4	80.7	360.2	333.5	348.7	310.6	369.7	512.0	435.6	422.8	346.0	374.8	314.8	211.9
		-	(0.3)	(0.3)	(0.3)	(0.3)	(0.2)	(0.4)	(0.3)	(0.3)	(0.3)	(0.7)	(0.3)	(0.4)	(0.3)	(0.4)
	AU	360.9	-	367.9	269.5	356.0	374.1	333.0	392.1	300.2	296.2	295.0	349.6	397.5	337.4	198.3
		(0.4)	-	(0.3)	(0.1)	(0.0)	(0.1)	(0.1)	(0.0)	(0.8)	(0.0)	(0.1)	(0.5)	(0.1)	(0.0)	(0.6)
	BR	89.0	366.7	-	378.8	268.8	303.4	246.8	305.2	473.9	342.0	401.9	325.1	311.7	250.4	198.1
		(0.7)	(0.2)	-	(0.3)	(0.0)	(0.1)	(0.2)	(0.0)	(0.3)	(0.2)	(0.7)	(0.4)	(0.1)	(0.1)	(0.4)
	CN	364.0	269.7	384.3	-	199.6	308.5	185.1	233.4	434.0	98.0	530.2	306.8	238.6	178.7	191.1
		(0.4)	(0.1)	(0.3)	-	(0.3)	(0.2)	(0.4)	(0.4)	(3.7)	(0.4)	(5.1)	(0.6)	(0.3)	(0.4)	(0.3)
	DE	337.2	356.0	269.1	200.3	-	62.1	31.3	81.8	211.4	271.3	341.9	61.8	55.3	27.1	145.4
		(0.5)	(0.0)	(0.0)	(0.3)	-	(0.0)	(0.1)	(0.0)	(0.1)	(0.2)	(0.5)	(0.3)	(0.1)	(0.0)	(0.6)
	FI	350.7	376.7	304.4	307.9	62.1	-	76.9	120.2	268.6	308.6	341.6	55.7	97.9	65.4	186.3
		(0.3)	(0.1)	(0.1)	(0.2)	(0.0)	-	(0.1)	(0.0)	(0.2)	(0.1)	(0.1)	(0.3)	(0.1)	(0.0)	(0.2)
	FR	313.9	332.7	245.4	185.4	30.6	76.7	-	66.9	189.5	271.5	289.5	69.0	72.5	11.9	154.3
		(0.5)	(0.1)	(0.2)	(0.4)	(0.1)	(0.1)	-	(0.1)	(0.1)	(0.2)	(0.2)	(0.6)	(0.1)	(0.1)	(1.1)
	IL	372.7	392.3	305.3	233.9	81.9	120.2	67.1	-	247.6	316.2	365.1	125.2	128.0	60.4	198.4
		(0.5)	(0.0)	(0.0)	(0.4)	(0.0)	(0.0)	(0.1)	-	(0.1)	(0.2)	(0.5)	(0.6)	(0.1)	(0.0)	(0.2)
	IN	504.3	309.0	470.4	430.6	214.2	266.2	192.5	250.8	-	180.9	321.1	235.5	256.5	194.6	330.5
		(0.3)	(0.8)	(0.3)	(3.7)	(0.1)	(0.2)	(0.2)	(0.1)	-	(0.2)	(0.9)	(0.4)	(0.2)	(0.1)	(0.2)
	JP	436.2	296.2	341.6	97.9	271.1	311.0	271.9	315.9	176.3	-	37.0	301.4	306.8	261.0	155.2
		(0.4)	(0.0)	(0.3)	(0.4)	(0.2)	(0.2)	(0.2)	(0.2)	(0.1)	-	(0.0)	(0.6)	(0.2)	(0.2)	(0.7)
	KR	424.2	293.9	399.5	506.7	347.8	341.9	292.0	375.5	320.6	37.4	-	328.4	356.4	285.2	191.5
		(0.7)	(0.1)	(0.6)	(4.9)	(0.4)	(0.1)	(0.3)	(0.5)	(0.6)	(0.0)	-	(0.5)	(0.3)	(0.2)	(0.5)
	RU	349.0	345.7	324.5	301.0	59.7	54.2	67.5	122.6	233.5	299.5	328.7	-	114.9	59.1	182.4
		(0.4)	(0.3)	(0.3)	(0.4)	(0.2)	(0.2)	(0.4)	(0.4)	(0.3)	(0.3)	(0.4)	-	(0.4)	(0.4)	(0.6)
	TR	384.3	399.7	311.6	239.1	56.7	99.3	75.3	129.4	256.9	308.1	355.3	115.2	-	67.7	181.8
		(0.7)	(0.1)	(0.1)	(0.2)	(0.1)	(0.1)	(0.2)	(0.1)	(0.2)	(0.2)	(0.3)	(0.4)	-	(0.1)	(0.2)
	UK	318.4	337.4	250.5	178.6	27.1	65.4	12.0	60.4	193.0	261.4	276.1	61.9	66.5	-	144.1
		(0.5)	(0.0)	(0.0)	(0.4)	(0.0)	(0.0)	(0.1)	(0.0)	(0.1)	(0.2)	(0.1)	(0.6)	(0.1)	-	(0.5)
	US	218.0	197.9	199.3	190.4	144.8	188.8	152.2	198.3	325.5	154.1	193.2	185.7	180.4	143.6	-
		(0.4)	(0.1)	(0.4)	(0.1)	(0.1)	(0.1)	(0.2)	(0.2)	(0.1)	(0.1)	(0.3)	(0.6)	(0.2)	(0.2)	-

Table 3. Observed mean RTTs (in ms; with 95% confidence intervals) between clients in Darmstadt (Germany), which used different Internet connection types, and globally distributed target nodes (effective sample size $n' \geq 991$).

		Target Node														
		AR	AU	BR	CN	DE	FI	FR	IL	IN	JP	KR	RU	TR	UK	US
Conn.	WLAN	303.4	341.4	276.6	380.4	30.7	90.9	30.5	91.8	194.1	282.8	354.8	128.1	77.5	35.1	182.9
		(0.9)	(0.6)	(0.4)	(3.4)	(0.4)	(0.2)	(0.4)	(0.3)	(0.7)	(1.7)	(0.4)	(3.2)	(0.4)	(0.6)	(1.5)
	UMTS	761.4	810.2	424.2	1043.1	44.6	410.3	41.0	342.2	427.6	562.5	853.1	405.9	412.1	47.0	433.6
		(40.8)	(42.5)	(11.9)	(41.0)	(1.1)	(18.3)	(2.9)	(24.8)	(12.3)	(27.2)	(43.2)	(10.9)	(11.5)	(2.6)	(14.1)

For example, for the Western European target hosts, such as Germany (DE), France (FR), or the United Kingdom (UK), we observed relative moderate absolute increases of 10 ms to 15 ms, or less than 50% in relative terms. In contrast, geographically more remote locations, such as Turkey (TR) or Israel (IL), were found to provide an acceptable latency around 80 ms when using WLAN, while the mean RTTs increase to more than 300 ms in UMTS networks.

In summary, with respect to the second research question from Section 1, we conclude that the placement of data centers close to the potential customers is a key factor in providing cloud gaming services with adequate QoS properties. Furthermore, the work at hand shows that not only WLAN, but also UMTS cellular networks may allow for cloud-based gaming if the cloud data centers that provide the according services are placed in geographical proximity to the user.

4 Related Work

Chen et al. have, to the best of our knowledge, been the first to conduct empirical latency measurements of actual cloud gaming providers [8]. In their experiments, they regarded *OnLive*, a commercial provider, as well as, *StreamMyGame*, a free software tool that permits to set up a private video game stream. Chen et al. propose and implement a measurement tool which is based on similar conceptual ideas as GALAMETO.KOM. Most notably, the authors also trigger a certain action – in their case, the invocation of the in-game menu – and observe the appearance of the corresponding reaction based on color changes. In their experiments, they find streaming delays – which do not include the network latency – between 135 ms and 240 ms for OnLive and up to 500 ms for StreamMyGame. Thus, their results are in a similar order of magnitude as the values that have been observed in our experiments. In contrast to this work, Chen et al. trigger the comparison process in the measurement tool through a redirected Direct3D function call and operate on the backbuffer of the graphics card, not the frontbuffer. Thus, the latency component that is introduced through the copying of the backbuffer scene into the frontbuffer has not been considered in their work. In addition, and more importantly, the authors do not use a locally executed game as benchmark in their experiments.

Jarschel et al. have conducted a user-study involving 58 participants on QoE of cloud gaming depending on network characteristics [4]. For that purpose, they generate an audio/video stream using a *PlayStation 3* gaming console. This stream is subjected to artificial delay and packet loss, ranging between 0 and 300 ms and 0 and 1% respectively, in different test scenarios. Jarschel et al. find that the quality of the downstream, i. e., the link between provider and user, has a substantially higher impact on the QoE than the quality of the upstream, i. e., the link between user and provider. Their results also indicate that packet loss is of higher relevance than latency for the subjective quality perception. The main difference compared to our work consists in the focus on subjective, rather than objective quality aspects. In addition, Jarschel et al. did not regard commercial cloud providers in their experiments.

Wang and Dey have proposed a cloud gaming system for mobile clients called *Cloud Mobile Gaming* (CMG) [25]. As part of their work, they examine the impact of different factors on the user experience. The considered factors involve the video stream configuration and quality, the game configuration, delay (i. e., latency), and packet loss. Similarly to Jarschel et al., the authors use a controlled experimental setup, in which they systematically vary the values of the previously mentioned factors. Using a study group of 21 participants, they infer impairment functions for these factors. The findings are subsequently validated using a control group of 15 participants. Based on practical measurements, the authors conclude that their CMG system may provide a subjectively good or mediocre gaming experience in WLAN and cellular networks, respectively. In contrast to our work, which considers public cloud gaming providers and the local execution of games, Wang and Dey exclusively examine their own, proprietary cloud gaming system.

Outside the academic world, West has measured the latency of various locally executed games on a PlayStation 3 console [26]. West uses a commodity digital camera in order to take high-frequency photographs of the game controller and the attached screen during gameplay. Based on a subsequent manual analysis of the resulting picture stream, he deduces the timespan between a button press and the corresponding action. West finds latencies between approximately 50 ms and 300 ms on the PlayStation 3. The main benefit ob West's method is the clear separation between the gaming system and the measurement system. In addition, the camera-based approach also permits to capture the LCD response time. However, the accuracy of the measurement is limited by the maximal framerate of the camera. In addition, GALAMETO.KOM only requires a brief preparatory manual tuning phase, whereas West's method requires substantial manual effort, which renders the collection of large data samples difficult.

Continuous measurements of latencies among various geographically distributed nodes are carried out within the so-called *Ping End-to-end Reporting* (PingER)[10] project. The project – which was started in 1995 and involves more than 700 worldwide nodes today – publicly provides its results in various formats through its Web site. Unfortunately, the project focuses on bilateral measurements and does not provide data on different network connections, such as cellular and wireless networks.

An open-source cloud gaming system, called *GamingAnywhere*, has been proposed by Huang et al. [13]. The software is available for public download through the project's Web site[11]. Huang et al. examine the performance of their system with respect to QoS and QoE properties through an experimental evaluation, based on three different games, and compare it to the performance of a commercial cloud gaming provider, namely OnLive, and a similar system, namely StreamMyGame. The authors find that their system provides latencies of around 40 ms to 45 ms, compared to about 150 ms to 200 ms for OnLive and approximately 400 ms for StreamMyGame. Based on their experiments, Huang et al. also claim that GamingAnywhere incurs substantially lower network load and features higher video quality than the two other systems. In contrast to our work, their paper does not include a comparison with local gaming and does not feature RTT measurements for globally distributed data centers.

In our our recent work [12], we have identified latency, energy consumption, and cost as main challenges for *mobile* cloud gaming. Similar to the work at hand, this previous publication provided unilateral RTT measurements in cellular and wireless networks. However, it did not feature large-scale bilateral measurements between globally distributed compute nodes, and did not consider actual cloud gaming providers.

Lastly, this invited paper builds on and extends our own previous publication [27]. As major novel contribution, the work at hand features bilateral and unilateral RTT measurements, which are a valuable complement to the primary experiment that focused on latency in cloud-based and local gaming.

[10] `http://www-iepm.slac.stanford.edu/pinger/`
[11] `http://gaminganywhere.org/`

5 Summary and Conclusions

The cloud computing paradigm has substantially transformed the delivery of IT services. A relatively new service class within this context is cloud gaming. In cloud gaming, video games are centrally executed in a cloud data center and delivered to the customer as an audio/video stream via the Internet. While this model has many advantages both from a user and provider perspective, it also introduces the Internet into the delivery chain, which may inflict the Quality of Service for the user.

In this work, our first focus was on the experimental evaluation of user-perceived latency in cloud-based and locally executed video games. For that matter, we created the semi-automatic measurement tool GALAMETO.KOM. We conducted latency measurement for two cloud gaming providers, using three different games and network types, respectively. Our results indicate that cloud gaming exhibits significantly higher latency than a local execution. Absolute increases were in the range between 40 ms and 150 ms, while the relative increases approximately amounted to between 85% and 800%. The margin between cloud providers and the local execution diminished with an improved network connection and an increase in computational complexity of the game.

As a complement to our primary experiment, this work featured an assessment of round-trip times among globally distributed compute nodes, as well as between these nodes and a single client that used different network connections. Here, we found mean round-trip times between 10 ms and 530 ms and 40 ms and 1050 ms for the first and second setup, respectively. These results confirm the notion that the provision of cloud games with adequate QoS properties require a placement of cloud data centers in geographical proximity to end users, specifically if cellular networks are used as delivery medium.

References

1. Dikaiakos, M., Katsaros, D., Mehra, P., Pallis, G., Vakali, A.: Cloud Computing: Distributed Internet Computing for IT and Scientific Research. IEEE Internet Computing **13**(5) (2009) 10–13
2. Buyya, R., Yeo, C., Venugopal, S., Broberg, J., Brandic, I.: Cloud Computing and Emerging IT Platforms: Vision, Hype, and Reality for Delivering Computing as the 5th Utility. Future Generation Computer Systems **25**(6) (2009) 599–616
3. Choy, S., Wong, B., Simon, G., Rosenberg, C.: The Brewing Storm in Cloud Gaming: A Measurement Study on Cloud to End-User Latency. In: 11th Annual Workshop on Network and Systems Support for Games. (2012)
4. Jarschel, M., Schlosser, D., Scheuring, S., Hossfeld, T.: An Evaluation of QoE in Cloud Gaming Based on Subjective Tests. In: 5th Int. Conf. on Innovative Mobile and Internet Services in Ubiquitous Computing. (2011)
5. Ross, P.: Cloud Computing's Killer App: Gaming. IEEE Spectrum **46**(3) (2009) 14
6. Süselbeck, R., Schiele, G., Becker, C.: Peer-to-Peer Support for Low-Latency Massively Multiplayer Online Games in the Cloud. In: 8th Annual Workshop on Network and Systems Support for Games. (2009)

7. Mell, P., Grance, T.: The NIST Definition of Cloud Computing – Special Publication 800-145. Technical report, National Institute of Standards and Technology (2011)
8. Chen, K., Chang, Y., Tseng, P., Huang, C., Lei, C.: Measuring the Latency of Cloud Gaming Systems. In: 19th ACM Int. Conf. on Multimedia. (2011)
9. Courcoubetis, C., Dramitinos, M., Stamoulis, G., Blocq, G., Miron, A., Orda, A.: Inter-Carrier Interconnection Services: QoS, Economics and Business Issues. In: 2011 IEEE Symposium on Computers and Communications. (2011)
10. Dick, M., Wellnitz, O., Wolf, L.: Analysis of Factors Affecting Players' Performance and Perception in Multiplayer Games. In: 4th ACM SIGCOMM Workshop on Network and System Support for Games. (2005)
11. Claypool, M., Claypool, K.: Latency Can Kill: Precision and Deadline in Online Games. In: First Annual ACM SIGMM Conf. on Multimedia Systems. (2010)
12. Lampe, U., Hans, R., Steinmetz, R.: Will Mobile Cloud Gaming Work? Findings on Latency, Energy, and Cost. In: 2nd Int. Conf. on Mobile Services. (2013)
13. Huang, C.Y., Hsu, C.H., Chang, Y.C., Chen, K.T.: GamingAnywhere: An Open Cloud Gaming System. In: 4th Multimedia Systems Conf. (2013)
14. Wang, J.: NVIDIA GeForce GRID – A Glimpse at the Future of Gaming. `http://www.geforce.com/whats-new/articles/geforce-grid` (May 2012)
15. Wilson, D.: Exploring Input Lag Inside and Out. `http://www.anandtech.com/show/2803/7` (Jul 2009)
16. Gaikai: Gaikai.com :: History. `http://www.gaikai.com/history` (Dec 2013)
17. Orland, K.: Report: Gaikai Streaming Coming to PS4 in Third Quarter of 2014 — Ars Technica. `http://arstechnica.com/gaming/2013/12/report-gaikai-streaming-coming-to-ps4-in-third-quarter-of-2014/` (Dec 2013)
18. Jain, R.K.: The Art of Computer Systems Performance Analysis: Techniques for Experimental Design, Measurement, Simulation, and Modeling. Wiley (1991)
19. Kirk, R.: Statistics: An Introduction. 5th edn. Wadsworth Publishing (2007)
20. Cai, W., Leung, V.: Multiplayer Cloud Gaming System with Cooperative Video Sharing. In: 4th Int. Conf. on Cloud Computing Technology and Science. (2012)
21. Tanenbaum, A.S.: Computer Networks. 4th edn. Pearson Education (2003)
22. Xiao, X.: Technical, Commercial and Regulatory Challenges of QoS: An Internet Service Model Perspective. Morgan Kaufmann (2008)
23. Greenberg, A., Hamilton, J., Maltz, D., Patel, P.: The Cost of a Cloud: Research Problems in Data Center Networks. ACM SIGCOMM Computer Communication Review **39**(1) (2008) 68–73
24. Chun, B., Culler, D., Roscoe, T., Bavier, A., Peterson, L., Wawrzoniak, M., Bowman, M.: PlanetLab: An Overlay Testbed for Broad-coverage Services. ACM SIGCOMM Computer Communication Review **33**(3) (2003) 3–12
25. Wang, S., Dey, S.: Modeling and Characterizing User Experience in a Cloud Server Based Mobile Gaming Approach. In: 2009 IEEE Global Telecommunications Conf. (2009)
26. West, M.: Measuring Responsiveness in Video Games. `http://www.gamasutra.com/view/feature/132122/measuring_responsiveness_in_video_.php` (Jul 2008)
27. Lampe, U., Wu, Q., Hans, R., Miede, A., Steinmetz, R.: To Frag Or To Be Fragged – An Empirical Assessment of Latency in Cloud Gaming. In: 3rd Int. Conf. on Cloud Computing and Services Science. (2013)

All online references in this paper were last accessed in January 2014.