Towards Cloud-Based Distributed Interactive Applications: Measurement, Modeling, and Analysis

Haiyang Wang, Tong Li*, Ryan Shea, Xiaoqiang Ma, Feng Wang, Jiangchuan Liu, *Fellow, IEEE*, Ke Xu*, *Senior Member, IEEE*

***Abstract*—With the prevalence of broadband network and wireless mobile network accesses, distributed interactive applications (DIAs) such as online gaming have attracted a vast number of users over the Internet. The deployment of these systems, however, comes with peculiar hardware/software requirements on the user consoles. Recently, such industrial pioneers as Gaikai, Onlive and Ciinow have offered a new generation of *cloud-based distributed interactive applications* (CDIAs), which shifts the necessary computing loads to cloud platforms and largely relieves the pressure on individual user's consoles.**

In this paper, we aim to understand the existing CDIA framework and highlight its design challenges. Our measurement reveals the inside structures as well as the operations of real CDIA systems and identifies the critical role of cloud proxies. While its design makes effective use of cloud resources to mitigate client's workloads, it may also significantly increase the interaction latency among clients if not carefully handled. Besides the extra network latency caused by cloud proxy involvement, we find that computation-intensive tasks (e.g., game video encoding) and bandwidth-intensive tasks (e.g., streaming the game screens to clients) together create a severe bottleneck in CDIA. Our experiment indicates that when the cloud proxies are virtual machines (VMs) in the cloud, the computation-intensive and bandwidth-intensive tasks may seriously interfere with each other. We accordingly capture this feature in our model and present an interference-aware solution. This solution not only smartly allocates workloads but also dynamically assigns capacities across VMs based on their arrival/departure patterns.

***Index Terms*—Cloud-based Distributed Interactive Application, Interaction Latency, Task Interference.**

I. Introduction

With unmatched system flexibility, distributed interactive applications (DIAs) have become increasingly popular in recent years. By providing diverse interactions among users, such applications as massive multiplayer online gaming, live messaging, and shared whiteboard have attracted a vast number of users over the Internet. Take online gaming as an example. It is reported in [1] that nowadays each US household on average owns at least one dedicated game console or PC for game playing, where 62% of them have played interactive games with others. The global markets of these DIA systems will also expand from 58.7 billion in 2011 to 83 billion in 2016, growing at a 7.2% compound annual rate [2]. Yet, to support superior interactions, DIAs often have peculiar demands on user consoles. The specialized consoles with high-performance hardware unavoidably increase user cost and greatly limit the penetration of DIAs to ubiquitous end users.

To bring true *play-as-you-go* into reality, industrial pioneers such as Gaikai [3], Onlive [4] and Ciinow [5] have suggested a new generation of DIAs based on cloud computing platforms. This *cloud-based distributed interactive application* (CDIA) effectively shifts the hardware/software requirements as well as the necessary computing loads to *cloud proxies*, and thus has attracted an increasing amount of attention form both service providers and end users. In particular, Sony Computer Entertainment (SCE) acquired Gaikai for 380 million USD on July 2, 2012, putting Gaikai as a key function in its next generation game console, PlayStation 4 [6]. Its competitor, Microsoft, also announced that the Gaikai-like CDIA functions will also play a major role in their new game console Xbox One [7]. Moreover, such industry leaders as AMD and Nvidia are also entering the market of CDIA services [8]. AMD's investment in CiiNow gives it a means of competing with rival Nvidia's GeForce Grid cloud gaming platform. In addition to cloud gaming, CDIAs can also take the form of the live notebook environments (e.g., Jupyter [9]) or a dashboard with live interactive graphical widgets (e.g., web applications based on R's shiny package [10]). These CDIAs also shift the software requirements to the cloud, and offload the compiling and parsing workloads to high-performance machines, i.e., cloud proxies.

Nowadays, CDIA remains in its infancy with plenty of unknown issues. In this paper, we take a first step towards understanding the CDIA framework and highlight its design challenges. Our measurement reveals the inside structure as well as the operations of real CDIA systems and identifies the critical role of cloud proxies. While this design makes effective use of cloud resources to mitigate client's workloads, it may also significantly increase the interaction latency among clients if not carefully handled. In both DIA and CDIA systems, the interaction latency is the most essential problem even when there is no limitation on the availability of server capacities [11] [12]. The increasing latency will unavoidably reduce the applicability of CDIA systems.

In detail, our analysis reveals that the deployment of cloud proxies adds extra communication hops between clients. To make the matter worse, the processing latency at the cloud proxy is also surprisingly high. While the use of the high-

Haiyang Wang is with the University of Minnesota at Duluth, USA (e-mail: haiyang@d.umn.edu). Tong Li is with Huawei Technologies Co., Ltd., China. (e-mail: litong12@tsinghua.org.cn). Ryan Shea and Jiangchuan Liu are with the Simon Fraser University, Canada (e-mail: {rws1, jcliu}@cs.sfu.ca). Xiaoqiang Ma is with the Huazhong University of Science and Technology, China (e-mail: mxqhust@gmail.com). Feng Wang is with the University of Mississippi, USA (e-mail: fwang@cs.olemiss.edu). Ke Xu is with the Tsinghua University, China. (e-mail: xuke@tsinghua.edu.cn).

*Ke Xu and Tong Li are the corresponding authors (e-mail: xuke@tsinghua.edu.cn, litong12@tsinghua.org.cn).

performance cloud platforms is expected to be highly efficient, we find that the computation-intensive tasks (e.g., game video encoding) and the bandwidth-intensive tasks (e.g., streaming game screens to clients) together create a severe bottleneck in CDIAs. Our experiment indicates that when the cloud proxies are virtual machines (VMs) in the cloud, the computation-intensive and bandwidth-intensive tasks may seriously interfere with each other if not handled carefully. An increase of traffic load will greatly slow down the CPU benchmark of cloud VMs. In the case of CDIA, when the cloud proxies are used to stream game screens to users, the computation-intensive operations, such as game processing and message forwarding, will also be invoked and prolong the interaction latency. The large number of CDIA users further aggravates this issue with mutual-interference, leading to poor user experiences. For example, *Diablo 2* in USEast realm has over 5 million users across 20 servers. The maximum acceptable latency of this Role Playing Games (RPG) is from 100 to 500 ms [13]. However, when we deploy these systems on CDIA, their latency can easily exceed 600 ms even when we assume that the network latency is as small as zero.

Such interference however does not exist in conventional physical machines or to a much lower degree. Therefore, the existing load assignment solutions in the DIA system are mainly focused on the optimization of stand-alone workloads, without considering their interference in the VM environment. To address this problem, we provide a model and consider an interference-aware solution that not only smartly allocates workloads but also dynamically assigns capacities across different VMs. Moreover, this model is also enhanced to capture user's arrival/departure patterns. The evaluation indicates that the proposed enhancement can successfully handle user dynamic while providing a bounded interaction latency.

The rest of this paper is organized as follows. Based on the measurement of Section II, we examine the latency issues in CDIA in Section III. Section IV reveals the task interference on cloud proxies and Section V provides a model to minimize the interaction latency. Our model is then extensively examined in Section VI. Section VII provides the solution to consider user's arrival/departure dynamic. Section VIII presents the related work and Section IX concludes the paper.

II. Cloud-Based DIA: Background and Framework

From the industrial perspective, CDIAs can bring immense benefits by expanding user bases to the vast number of less-powerful devices that support thin clients only, particularly smartphones and tablets. As an example, the recommended system configuration for *Battlefield 3*, a highly popular first-person shooter game, is a quad-core CPU, 4 GB RAM, 20 GB storage space, and a graphics card with at least 1GB RAM (e.g., NVIDIA GEFORCE GTX 560 or ATI RADEON 6950), which alone costs more than $500 USD. The newest tablets (e.g., Apple's iPad with Retina display and Google's Nexus 10) cannot even meet the minimum system requirements that need a dual-core CPU over 2.4 GHz, 2 GB RAM, and a graphics card with 512 MB RAM, not to mention smartphones of which the hardware is limited by their smaller size and

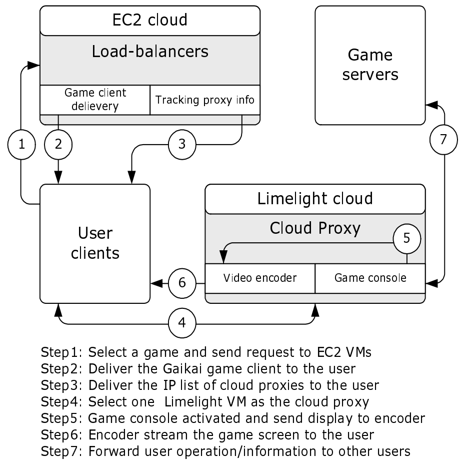

EC2 cloud
Load-balancers
Game client delievery
Tracking proxy info
Game servers
1
2
3
7
User clients
Limelight cloud
Cloud Proxy
5
Video encoder
Game console
6
4
Step1: Select a game and send request to EC2 VMs
Step2: Deliver the Gaikai game client to the user
Step3: Deliver the IP list of cloud proxies to the user
Step4: Select one Limelight VM as the cloud proxy
Step5: Game console activated and send display to encoder
Step6: Encoder stream the game screen to the user
Step7: Forward user operation/information to other users

Fig. 1: Basic Framework of Gaikai

thermal control. Furthermore, mobile terminals have different hardware/software architectures from PCs, e.g., ARM rather than x86 for CPU, lower memory frequency and bandwidth, power limitations, and distinct operating systems. Therefore, the traditional console game model is unfeasible for these devices, which in turn become targets for such CDIA systems as Gaikai and Onlive. Different from existing DIAs, CDIAs utilize the powerful and elastic service capacity offered by cloud computing to mitigate the hardware/software requirements on user consoles. In particular, Gaikai and Onlive deploy the game clients/consoles on cloud platforms and only stream the game screen/interactions to end users.

To understand how CDIAs work in detail, we take Gaikai as a case study. Since 2011, it has emerged as one of the most popular cloud-based online gaming systems with over 100 million subscribers. It not only provides free PC game demos but also powers high quality gaming experiences onto smartphones, tablets and Internet TVs [14].

Our experiments have conducted traffic measurement and analysis from the edge of four networks, which are located in four different countries (United States, Canada, China and Japan) in two distinct continents. We used traffic analysis, shared library and system call interception techniques to analyze various aspects of the Gaikai protocol. We also applies a packet level analysis to understand the communications between our clients and cloud proxies. In particular, we monitor Gaikai's online gaming service with clients from different network locations and capture their traffic from/to the Gaikai servers. We first examine the details in the control messages, look into the packet payloads and identify the information, such as the domain names, of Gaikai's basic components. After that, we carefully analyze the data traffic to further understand the functions of these components. To avoid possible bias, we apply classic analysis tools (e.g., ISP lookup) that are widely adopted in the existing reverse engineering studies [15]. Note that some online games are designed to utilize one of the user clients as the server to enable interactions. We therefore take advantage of this feature and capture the server-side traffic to better understand the related protocols.

Through analyzing the captured traffic, we illustrate

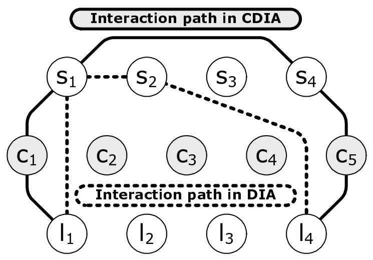


Fig. 2: Path of client interaction

Gaikai's basic framework/protocol in Figure 1. We can see that there are two major components on the server side of Gaikai (marked as grey boxes in the figure). The first part is Amazon EC2-based [16] load-balancers[1], and the second part is the Limelight-based game proxy servers [17]. Both Amazon and Limelight are leading cloud service providers with Xen virtualization [18]. Gaikai applies both platforms to accomplish different functionalities and utilizes their widely geo-distributed instances to push these functions closer to the users.

When a user selects a game on Gaikai (*Step 1* in Figure 1), an EC2 virtual machine (VM) will first deliver the Gaikai game client to the user (in *Step 2*). After that, it forwards the IP addresses of the available Limelight game proxies to the users (in *Step 3*). The user will then use one game proxy to run the game (in *Step 4*). To ensure smooth game playing, this selected game proxy uses a *packet train measurement* [19] to estimate the available bandwidth to the users. This is identified by our packet-level analysis, which shows that the game proxy sends back-to-back packets with empty payload to test the available bandwidth. Note that the game proxy starts the game only when the available bandwidth can well-support an FPS (frames per second) around 60 for video streaming. After that, the game proxy starts to run the game and the game screen will be streamed back to the user via UDP (in *Step 5* and *Step 6*). For multiplayer online games, these game proxies will also forward user operations to the game servers (mostly deployed by game developers) and send the related information/reactions back to users (in *Step 7*). It is easy to see that such a CDIA system can remarkably relieve the hardware/software requirements on the user side, given that now the games are running on the cloud platforms. This change enables users to play hard-core games over much less powerful devices, e.g., over smartphone, tablets, or even digital TVs, as long as they are multimedia- and network-ready.

We have also measured other CDIA platforms, and have found that Gaikai's framework is representative, which is not surprising given it as a very natural extension to the conventional DIA with cloud assistance.

III. Interaction Latency of CDIA: Issues and Challenges

The CDIA framework offers great opportunities for both users and service providers. Similar to the conventional DIA systems, its service performance is highly depending on users' *interaction latency* [12]. In particular, the CDIA systems, such as the online cloud gaming applications, need to collect user actions, transmit them to the cloud proxy, process the action, render the results, encode/compress the resulting changes to the game-world, and stream the video (game scenes) back to the player. To ensure interactivity, all of these serial operations must happen in the order of milliseconds. The interaction latency is thus essential even when there are no limitations on the availability of server capacities [11].

Figure 2 illustrates the interaction pathes in both DIA and CDIA frameworks, where L is the set of clients, S is the set of servers, and C is the set of cloud-based proxies. To better compare the performance of DIA and CIDA design, we will first focus on the latency between clients and servers (e.g., between l_1 and s_1). It is easy to see that such a latency consist of two parts in the CDIA systems: First, the network latency between clients and servers; Second, the processing latency on cloud proxies. We will provide a step-by-step discussion to understand these two parts in the following subsections.

A. Network Latency Analysis

To clarify the extra network latency in CDIA design, we carry out a real-word experiment from Planet-lab. We use a server in our campus to emulate the game server in CDIA[2]. We select 588 Planet-lab nodes (the maximum number of nodes that we can access) to run as CDIA clients and emulate the CDIA framework by using the server and these clients to connect Gaikai's cloud proxies. We have found 28 cloud proxies during the measurement of Gaikai[3], and therefore use the IP addresses of these proxies in this experiment. We first measure the RTTs between 588 Planet-lab clients and 28 Gaikai cloud proxies and then the RTTs between the server and these cloud proxies. The sum of these two latencies can be used to calculate the client-server RTTs in this CDIA system. To provide a fair comparison, we also measure the direct RTTs between the server and the Planet-lab clients as the baseline (the case of conventional DIA).

Figure 3 compares the client-server RTT in both DIA and CDIA. We can see that most (over 80%) users in DIA have quite low client-server latency (less than 60 ms), while the average latency is much worse if we put them into CDIA, as shown in Figure 4. The worst case in Figure 4 shows 90% users have an interaction latency over 200 ms. This is hardly acceptable for smooth interaction because the processing latencies are not yet considered in this experiment. It is known that adding extra nodes in any overlay network may unnecessarily lead to longer path length given that triangle inequality does not hold in the Internet [20]. Hence, there is indeed space to reduce the latency beyond naive proxy deployment[4].

[1]Based on our measurement, they also have other functions beside load-balancing. We call them *load-balancers* because Gaikai marks them with "LB" in their domain names.

[2]We have observed similar results over 50 servers that are located in different places.

[3]The total number of Gaikai's cloud proxy is unknown to the general public. These sampled cloud proxies are used to estimate the network latency in such a system.

[4]The 588 PlanetLab nodes are applied in both DIA and CDIA experiments to provide fair comparison. Since some real-world interactive applications, such as Diablo 2, may divide their users into realms, we also investigate the case using a subset of PlanetLab nodes from one realm (e.g., USEast). The results remain consistent to Figure 3.

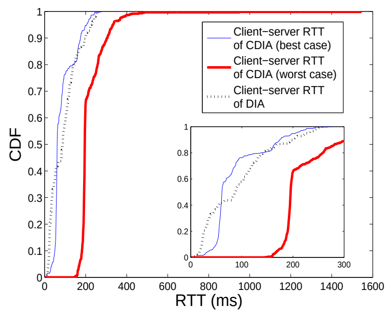


Fig. 3: Time cost between user and server (DIA v.s. CDIA)

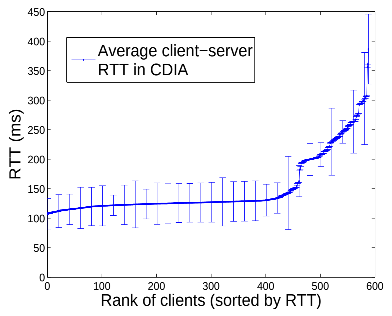


Fig. 4: Average user-server latency in CDIA

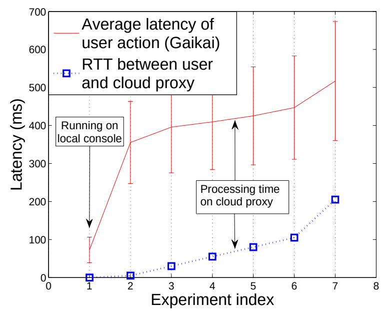


Fig. 5: Average latency of user's action

B. Processing Latency Analysis

So far, we have only considered the network latency in the CDIA framework. It is easy to see that the cloud proxies will also bring extra processing latency to the interaction. We now closely examine this latency and identify its impacts.

To focus exactly on the interaction between clients and cloud proxies, we select a single player game where a player (client) does not need to communicate with the game server and other players. The player simply sends the operations to the cloud proxy and the proxy then streams the responding game screen back to the player. Since the RTT between the player and the cloud proxy can be directly measured, we only need to obtain the *response time*[5], which, after subtracting the RTT, gives the processing latency at the cloud proxy. The detail of this experiment is as follows.

We first select an action button in the game *The Witcher 2: Assassins of Kings*; in particular, the "map" button. We click this button and start to record the game screen at 100 FPS (frames per second). This sampling rate already exceeds the normal game play which is around 60 to 70 FPS. We then check the video file frame-by-frame until we find the frame where the map is displayed. We run this experiment for multiple times under different RTTs. These RTTs are controlled by the traffic shaping tool *TC* [21]. To better understand the processing overhead at the cloud proxy, we also record the response time on a local game console. We use the same game on both Gaikai and the local console to provide a fair comparison.

TABLE I: Delay Tolerance in Traditional Gaming

Example Game Type	Perspective	Delay Threshold
First Person Shooter (FPS)	First Person	100 ms
Role Playing Game (RPG)	Third-Person	500 ms
Real Time Strategy (RTS)	Omnipresent	1000 ms

As we can see from Figure 5, the local console general needs 80 ms to open the map for the players with very small standard deviation. Note that the RTT is zero in this case because the game is locally rendered. When we run this game remotely on Gaikai, the response time elevates to more than 300 ms. The overhead (in terms of the processing latency) on the Gaikai proxy is thus approximately 220 ms[6]. When we consider the interactive latency between different users, there will be, unfortunately, two proxies in their interactive path. This means the interactive latency can easily exceed 600 ms. It is worth noting that the studies on traditional gaming systems have found that different styles of games have different thresholds for maximum tolerable delays [13], [22]–[24]. These thresholds can be further adjusted according to the user feedback and operating experience, this is because that indeed the slower the game is, the less delay influences the user rating [25]. However, latency as high as 600 ms will not be acceptable for most online gaming applications as shown in Table I.

IV. Interferences between Computation-intensive and Bandwidth-intensive Tasks

It is surprising to see that the cloud proxies can introduce such a high processing latency in CDIA. This is unlikely caused by video encoding only since many CDIA service providers have claimed that their video encoding latency is indeed very small within 10 ms. We therefore further explore its underlying reasons in this subsection.

The cloud proxy on Gaikai is different from a local game console. It is a virtual machine (VM) running both computation-intensive tasks (for example, rendering the game) and bandwidth-intensive tasks (for example, streaming the game screen to the players) at the same time. Since these tasks cannot be decoupled into different VMs, it is necessary to see if they introduce extra overheads on the cloud proxy. Note that currently most cloud service providers usually provide the instances with custom optimization for DIA's specific needs [26] as well as the DIA's completely control over the instances [27], [28]. In other words, DIAs are usually assigned dedicated cloud instances. Thus, in this paper, we do not consider the interferences among different concurrent DIAs on the same instance.

Since the Limelight platform can hardly be tested by individual users, we run the standard CPU benchmark created by a tool called *sysbench* [29] on EC2 instances (their cloud platforms are quite similar in terms of Xen-based cloud virtualization). We use network tool *iPerf* [30] to adjust the

[5]The response time is the latency that the player waits until the result of her/his operations is returned. For example, if the player clicks the button "option" at time t_i and the option menu displays at time t_j, the response time can be calculated as $t_j - t_i$.

[6]To avoid measurement bias, we also test actions that make different changes to the in game world, for example, small character movements. The results remain consistent with Figure 5.

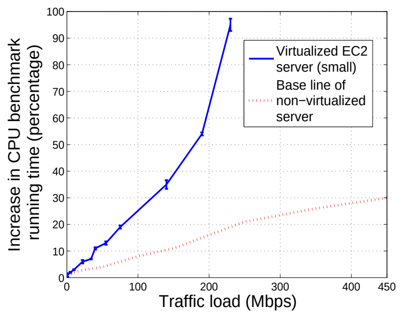


Fig. 6: Increase of CPU benchmark running time (on EC2 small instance)

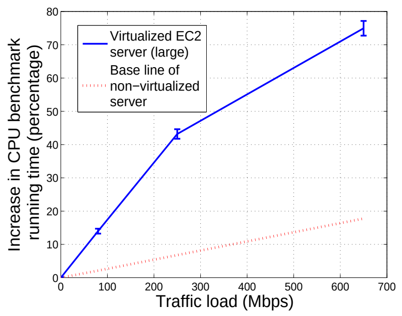


Fig. 7: Increase of CPU benchmark running time (on EC2 large instance)

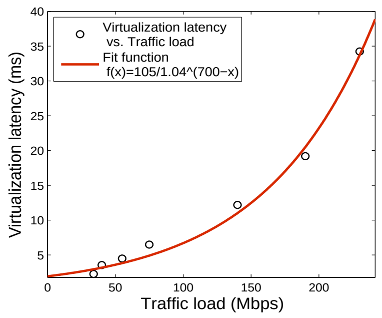


Fig. 8: Curve fitting for Equation (4) (on EC2 small instance)

traffic load on the instance and check the time cost of running the benchmark. To provide a fair comparison, we also do the experiment on local servers (non-virtualized servers) as a baseline. Figure 6 shows a comparison between an EC2 small instance and our local server. In this experiment, the EC2 small instance has 1.7 GB memory, 1 EC2 compute unit (Intel Xeon series, 1 virtual core with 1 EC2 compute unit), and 160 GB instance storage with 32-bit platform. Our local server also has similar hardware configuration that is comparable to the EC2 small instance. From this figure, we can see that the traffic load on the non-virtualized server only slightly increases the running time of the CPU benchmark, e.g., 250 Mbps traffic load will only increase the running time by 20%. However, for the virtualized EC2 small instance, this traffic load will double the running time of the CPU benchmark with very small standard deviation (the detailed data can be found in Table II). We have also tested this on EC2 large instances with multiple cores. The large instance has 7.5 GB memory, 4 EC2 computation units (Intel Xeon E5-2676 v3, 2 virtual cores and each with 2 EC2 computation units), 850 GB instance storage with 64-bit platform and very high I/O performance. Our local server, on the other hand, has weaker hardware configuration, particularly the CPU capacity. As shown in Figure 7, we can see that the traffic load on large instances still brings remarkable overheads to the system. Although the result is better than that of small instances, the traffic load will still remarkably slow down the running time of the CPU benchmark especially when comparing with the non-virtualized baseline.

It is easy to see that the CDIA cloud proxies are indeed in the same situation as these EC2 instances. The traffic load can significantly slow down the game running and unavoidably leads to a high processing latency. Yet, such a problem is rarely seen on the non-virtualized local game consoles or cloud proxies, or to a much lower degree.

V. Latency Minimization in CDIA

Given the importance of interaction latency, there have been significant studies on latency minimization for conventional DIAs, mostly focusing on latency directly between client pairs [31] [32] [12]. Unfortunately, the existence of cloud proxies prevents them from being used in CDIAs, not to menton the task interference on the cloud proxies. In this section, we will revisit the latency modeling problem in these new contexts. We first consider a basic model to optimize the network latency. After that, this model will be further enhanced to capture the task interference on the cloud VMs.

TABLE II: CPU benchmark running time under different traffic loads on virtualized EC2 small instance

Load	**Run 1**	**Run 2**	**Run 3**	**Run 4**	**Avg**
	Running time of CPU benchmark(ms)				
1 Mbps	37.17	36.71	36.75	36.64	36.82
5 Mbps	36.92	37.06	37.07	37.07	37.03
10 Mbps	37.41	37.51	37.36	37.42	37.42
22 Mbps	38.21	38.49	38.75	38.30	38.44
34 Mbps	39.03	39.13	39.06	39.13	39.09
40 Mbps	40.33	40.59	40.39	40.20	40.38
55 Mbps	41.28	41.22	41.06	41.67	41.31
75 Mbps	43.20	43.21	43.14	43.69	43.31
140 Mbps	49.00	48.84	50.06	48.13	49.01
190 Mbps	56.28	56.26	55.60	55.96	56.02
230 Mbps	71.50	70.67	69.06	73.01	71.06

A. *Basic Model of CDIA Latency*

To ensure responsive interactions, previous studies have suggested that reducing the average latency is not enough, because any fast users would suffer when they interact with long latency users [33] [34]. Our objective is thus to minimize the maximum latency between all client pairs that are bridged by cloud proxies. We focus on network latency here and will address processing latency in the next two sections.

$S = \{s_1, s_2, ..., s_m\}$ denotes the set of m servers and $L = \{l_1, l_2, ..., l_n\}$ denotes the set of n user clients. Let $C = \{c_1, c_2, ..., c_o\}$ be the set of o cloud proxies. Each client in L will be assigned to a cloud proxy and a server in order to send operations and receive updates from other clients. An assignment A is a mapping from L to C and S, where for each client $l \in L$, $c_A(l) \in C$ denotes the assigned cloud proxy of client l and $s_A(l) \in S$ denotes the assigned server of client l.

For clients l_i and l_j, the communication should go through their assigned cloud proxies and servers in CDIA. Specifically, if l_i issues an operation to l_j, the following steps should be taken so that l_j can see the effect of this operation. First, l_i sends the operation to its assigned cloud proxy $c_A(l_i)$. $c_A(l_i)$

will then forward this operation to server $s_A(l_i)$ that is also assigned to l_i; After that, if l_j is assigned to a different server $s_A(l_j)$, server $s_A(l_i)$ should forward the operation to server $s_A(l_j)$; Then $s_A(l_j)$ executes the operation and delivers the resultant state update to l_j's cloud proxy $c_A(l_j)$; Finally, $c_A(l_j)$ will generate the game screen and stream the display to client l_j. Let $D(u,v)$ be the path latency between two nodes that are not directly connected and $d(u,v)$ be the link latency between two neighbor nodes. To be consistent with the existing DIA models [12], we assume that $D(u,v)=D(v,u)$ and $d(u,v)=d(v,u)$. The latency between client l_i and its server $s_A(l_i)$ can be calculated as:

$$D\Big(l_i, s_A(l_i)\Big) = d\Big(l_i, c_A(l_i)\Big) + d\Big(c_A(l_i), s_A(l_i)\Big) \tag{1}$$

We can therefore obtain the total interaction latency between l_i and l_j as follows:

$$\begin{aligned} D(l_i,l_j) = D\Big(l_i, s_A(l_i)\Big) + D\Big(l_j, s_A(l_j)\Big) \\ +d\Big(s_A(l_i), s_A(l_j)\Big)\cdot I_{\big[s_A(l_i)\neq s_A(l_j)\big]} \end{aligned} \tag{2}$$

where $d\Big(s_A(l_i), s_A(l_j)\Big)$ denotes the latency between server $s_A(l_i)$ and $s_A(l_j)$, and $I_{[\cdot]}$ indicates whether l_i and l_j are assigned to different servers (1: yes; 0: no). Given the interaction latency between l_i and l_j, our objective is to find an assignment A to minimize $\mathbb{U}(A)$, the maximum interaction latency among all client pairs:

$$minimize \qquad \mathbb{U}(A) = \max_{l_i,l_j\in L}\Big\{D(l_i,l_j)\Big\} \tag{3}$$

This min-max latency can be optimally found when we convert it into a longest path problem in directed acyclic graph (DAG). The details are presented in *Appendix A*.

B. Enhanced Model to Capture Task Interference

In this part, we will further extend our model to consider the impact of traffic load on different cloud proxies. It is worth noting that CDIA offers elastic service capacity at cloud proxies. The cloud proxy capacity can be dynamically adjusted to meet user demands. Therefore, we denote P as the cloud proxy capacity set where $P=\{p_1,p_2,...,p_o\}$; $p_i\in P$ refers to the amount of resource that is assigned to cloud proxy c_i (bandwidth capacity in this case)[7]. Based on our measurement, we find that the NPV (Net Present Value) function [35] can be borrowed to capture the relationship between *virtualization latency* (processing latency caused by traffic load on VMs) and traffic load[8], we therefore compute the virtualization latency of cloud c_i as:

$$r(p_i) = \frac{a}{b^{p_i - q_A(c_i)}} \tag{4}$$

[7]We assume that one cloud proxy can serve multiple clients. Although using dedicated VMs reduces interference and improves the Quality of Experience (QoE), it dramatically increases the cost taking into account the large amount of users. Moreover, since we focus on the interference between computation-intensive tasks and bandwidth-intensive tasks, this interference still exists in the dedicated VM solution.

[8]This function has been widely used to quantify the relationship between cash and price/cost, which resembles our case when we try to purchase more cloud resources to reduce the virtualization cost on VMs.

where a indicates the latency when the cloud proxy is fully loaded (without remaining bandwidth). Parameter b controls the skewness of the relationship between load and latency where $b\in(1,+\infty)$. Note that different VMs may have different a and b. For example, in Figure 8, a is around 105 and b is around 1.04.

Given a load assignment A and user l_i, we denote $p_A(l_i)$ as the resource assigned to cloud proxy $c_A(l_i)$. For a given set of servers, $S=\{s_1,s_2,...,s_m\}$ and clients $L=\{l_1,l_2,...,l_n\}$, the problem becomes how to use a set of cloud proxies $C=\{c_1,c_2,...,c_o\}$ to connect these clients and servers, with load assignment A and resource assignment P, to minimize the maximum interaction latency between all client pairs:

$$\begin{aligned} minimize \qquad \mathbb{U}(A,P) = \underset{l_i,l_j\in L}{Max}\Big\{D(l_i,l_j) \\ +r(p_A(l_i)) + r(p_A(l_j))\Big\} \end{aligned} \tag{5}$$

$$s.t. \quad \forall i=1,2,...,o, \qquad q_A(c_i)\le p_i \tag{6}$$

$$\sum_{i=1}^{o} p_i * Cost(c_i) \le K \tag{7}$$

where K refers to the total budget, which we assume can at least serve all the clients in the system.

It is easy to see that the virtualization latency makes the problem harder. If we assign client l_i to $c_A(l_i)$, it will not only assign traffic load to $c_A(l_i)$ but also affect the performance of other clients who have also been assigned to this cloud proxy. Assuming the cloud proxy capacity is given, this client assignment problem can therefore be transformed into a $0-1$ Multiple Knapsack problem with a non-linear objective function, which is known to be NP-hard [36].

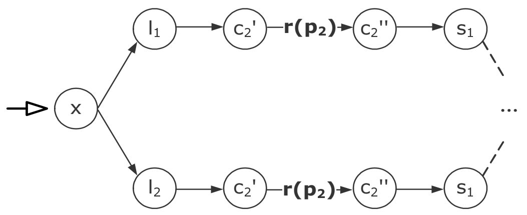


Fig. 9: Transform G into G_A^*

By exhaustively searching along all the possible combinations of A and P, the optimal solution can be achieved. However, the practical effectiveness of this search is limited considering the real-time user demands in CDIA systems. We thus propose a bi-level heuristic, which divides the optimization problem into two stages: load assignment and resource assignment. In the load assignment stage, we assume that all the cloud proxies are fully loaded (the virtualization latency is therefore equal to a in Equation (4)) and find the optimal load assignment A by Algorithm 5 (in *Appendix A*). After that, we construct a subgraph G_A based on the existing graph G and assignment A. As shown in Figure 9, we then split the node c_i into two virtual nodes (c_i' and c_i'') and use their link weight to refer to the virtualization cost on c_i. We denote G_A^* as the resulting graph and further apply a greedy algorithm to find the resource assignment P in G_A^*. As shown in Algorithm 1, this greedy algorithm iteratively assigns resource to the cloud

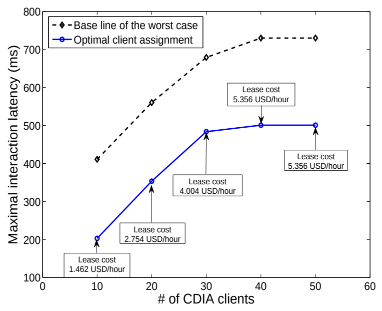


Fig. 10: Optimal client assignment only considering the networking latency

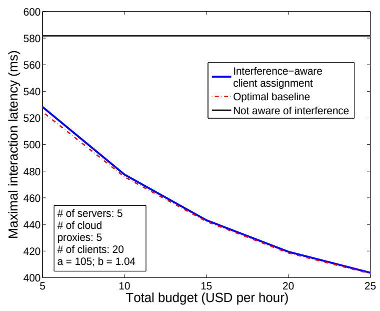


Fig. 11: Interference-aware client assignment

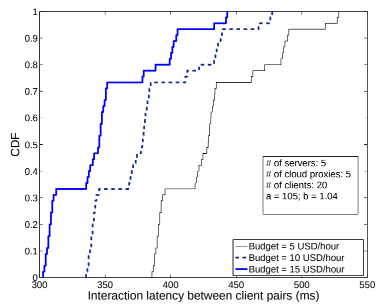


Fig. 12: Interaction latency across client pairs (with different budgets)

proxies on the longest path. The algorithm stops when the remaining budget is not enough. In the next section, we will show that this bi-level heuristic achieves near-optimal performance in practical settings.

Algorithm 1 ResourceProvisioning()

```
 1: Get $G_A^*$ from $A$;
 2: $R \leftarrow K - \mathbb{C}$;
 3: while (true) do
 4:     $path^* = LongestPath(G_A^*)$;
 5:     Get $c_i, c_j$ from $path^*$;
 6:     if $R \geq max(Cost(c_i), Cost(c_j))$ then
 7:         if $\frac{r(p_i)-r(p_i+1)}{Cost(c_i)} \geq \frac{r(p_j)-r(p_j+1)}{Cost(c_j)}$ then
 8:             $R \leftarrow R - Cost(c_i)$;
 9:             $p_i \leftarrow p_i + 1$;
10:         else
11:             $R \leftarrow R - Cost(c_j)$;
12:             $p_j \leftarrow p_j + 1$;
13:         end if;
14:     else if $R \geq min(Cost(c_i), Cost(c_j))$ then
15:         if $Cost(c_i) \leq Cost(c_j)$ then
16:             $w \leftarrow i$;
17:         else
18:             $w \leftarrow j$;
19:             $R \leftarrow R - Cost(c_w)$;
20:             $p(c_w) \leftarrow p(c_w) + 1$;
21:         end if;
22:     else
23:         break;
24:     end if;
25: end while;
```

VI. Performance Evaluation

We now evaluate the performance of our solution via extensive trace-based simulations in MATLAB. The network latency (measured in Section IV) and the processing delay (measured in Section VI) will both serve as the inputs of our evaluation. We first examine the performance of our optimal client assignment when there is no task interference[9]. After that, we investigate the performance of the interference-aware client assignment algorithm in the virtualized environment.

We start with a CDIA system that consists of 20 clients, 5 cloud proxies and 5 servers. The renting cost of cloud proxies are referenced from the instance price list of Amazon's *On Demand instances* [16]. Figure 10 presents the performance of our optimal client assignment when there is no task interference (the processing latency is a default value of 80 ms at the cloud proxies). It is easy to see that the smart client assignment greatly reduces the interaction latency. Without optimization, the maximum client interaction latency can be as high as 700 ms. Our approach, on the other hand, can reduce the maximum latency to less than 500 ms. It is also worth noting that the renting price is linearly related to the client population. This indicates a good scalability of our approach.

It is not surprising to see that the optimal client assignment can achieve such a significant gain when there is no task interference. Figure 11 further explores the case when the optimal assignment can hardly be archived in the task interference environment. We can see that task interference optimization is critical for CDIAs. The maximum interaction latency can be larger than 580 ms if we focus on it only. Fortunately, our interference-aware algorithm can achieve a near-optimal (with the difference within 5 ms) latency that greatly reduces the interaction latency[10]. It is worth noting that the interaction latency can be further reduced and become closer to the optimal results when we have more budget to purchase more capacities at the cloud proxies.

Figure 12 takes a closer look at the interaction latency between individual clients. We can see that all clients can benefit from the total budget increase. To be more specific, when the budget is equal to 5 USD/hour, less than 30% clients can have an interaction latency less than 400 ms. If we increase the budget to 15 USD/hour, more than 95% clients can interact with each other with a latency below 400 ms. The difference between the fastest and the slowest clients is also quite small, around 150 ms. Figure 13 further shows the cases with different number of CDIA clients. We can see that for a given budget, our algorithm scales well with an increasing number of clients. Note that the total budget also bounds the total capacity of the cloud proxies. We thus cannot add more clients in Figure 13.

[9]This will be the case when the system is deployed on non-virtualized cloud platforms.

[10]The optimal base-line is obtained by brute-force searching.

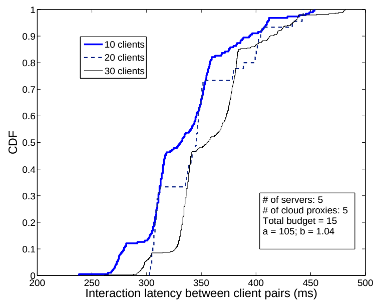


Fig. 13: Interaction latency across client pairs (different # of clients)

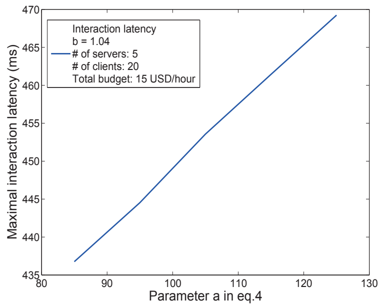


Fig. 14: Adjusting parameter a (VM's maximum processing latency)

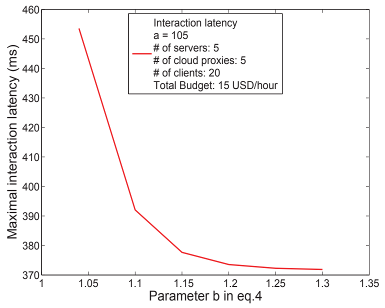


Fig. 15: Adjusting parameter b (skewness of the relationship)

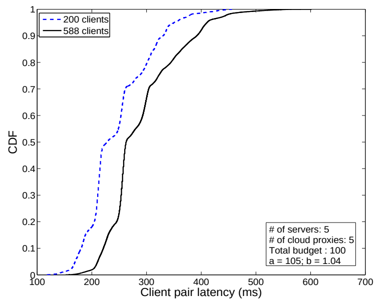


Fig. 16: Interaction latency across 200 and 588 clients

To understand the virtualization latency on different types of VMs, we investigate the case with different parameter inputs in Equation (4). Figure 14 presents the case when the maximum processing latency (parameter a for the cloud proxies) is changed from 85 ms to 125 ms. We can see that the interaction latency increases linearly with a. On the other hand, Figure 15 presents the case when virtualization latency and traffic load have more skewed relationships[11]. Based on these two figures, we can find that a good VM should have a small a and a large b. The maximum processing latency should be small when the VM is fully loaded (with a small a). In other words, adding idle resources on the VM ought to significantly reduce such a processing latency (with a large b).

Figure 16 further presents the CDF of the interaction latency across 200 random selected clients and all the 588 clients[12] in our measurement, respectively. It is easy to see that 80% clients can achieve the interaction latency within 300 ms. The interaction latencies between most (70%) client pairs are between 200 ms and 250 ms. It is also worth noting that the total budget in this case is relatively high up to 100 USD per hour. This is because we are using the pricing list of Amazon's *On Demand instances*. Choosing other types of platforms/instances, such as the *Reserved instance* may further reduce this cost.

VII. System Scaling with User Dynamic

The above model for the first time considers the task interference in classic CDIA systems. The related analysis can be applied to understand its budget-QoS relationship [37]. For example, how much money can archive what level of min-max interaction latency across how many users. In this section, we will further examine the system scaling issues with the arrival and the departure of clients. When a client arrives/departs the system, we need to minimize the maximum interaction latency with a bounded budget (in Equations (5), (6) and (7)). To better decide when we need to acquire/release extra cloud resource, we introduce a threshold ξ ($\xi \geq T_{max}$) to denote user's delay tolerance in different CDIA systems, where T_{max} is the maximum interaction latency across all clients. As we have discussed in Table I, the configuration of ξ is related to the type of different CDIA applications. This threshold adds the following new constraint to Equation (5):

$$\mathbb{U}(A, P) \leq \xi \tag{8}$$

Based on this new constraint, we then discuss our design to handle the client arrival/departure-aware step-by-step.

Algorithm 2 ClientArrival()

```
1: for all $c_i$ that $c_i \in C$ do
2:     Calculate the expected processing time $r^*(c_i)$ when $c(\triangle l) \leftarrow c_i$;
3:     $\triangle T(c_i) = \max\{r^*(c_i) + D(\triangle l, l_j)\}$, $l_j \in L$;
4: end for;
5: Get $\triangle T_{min} = \min\{\triangle T(c_i) | c_i \in C\}$;
6: if $\triangle T_{min} > \xi$ then
7:     OnlineResourceProvisioning();
8: else
9:     $\triangle c = \{c_i | \triangle T(c_i) = \triangle T_{min}, c_i \in C\}$;
10:    Calculate $T_{max}$ when $c(\triangle l) = \triangle c$;
11:    if $T_{max} > \xi$ then
12:        OnlineResourceProvisioning();
13:    else
14:        $c(\triangle l) \leftarrow \triangle c$;
15:        if $p(c(\triangle l)) = q(\triangle l)$ then
16:            $p(c(\triangle l)) = p(c(\triangle l)) + 1$;
17:        end if;
18:    end if;
19: end if;
20: $q(c(\triangle l)) = q(c(\triangle l)) + 1$;
21: return $c(\triangle l)$;
```

[11]Note that different a, b pairs in these two figures can present different cloud instances. For example, we use $a = 105$ and $b = 1.04$ to capture the features of EC2 small instances in our simulation.

[12]All the clients that we have used in our measurement in Section IV.

A. Handling Client Arrival and Departure

For any two clients $l_i, l_j \in L$, to emphasize the simultaneous consideration of communication and computation, we denote $T(l_i, l_j)$ as the interaction latency between client l_i and l_j, then we have $T_{max} = \max\{T(l_i, l_j)\}$. As to a new arriving client $\triangle l$, $\triangle T$ denotes the maximum interaction latency between $\triangle l$ and the other clients, i.e., $\triangle T = \max\{T(\triangle l, l_i)|l_i \in L\}$. Since $\triangle T$ depends on the mapping from C to $\triangle l$, we define the min-max interaction latency across all proxies as $\triangle T_{min} = \min\{\triangle T(c_i)|c_i \in C\}$.

Algorithm 3 OnlineResourceProvisioning()

```
1:  $\triangle p \leftarrow 1$, $target \leftarrow 0$, $flag \leftarrow 1$;
2:  Sort $C$ by ascendant order of $Cost(c_i)$;
3:  while ($flag$) do
4:    for all $c_i$ that $c_i \in C$ do
5:      $c(\triangle l) \leftarrow c_i$;
6:      Get the client $l'$ that has the maximal interaction
        delay with $\triangle l$, then get $c_j = c(l')$;
7:      if $CP(c_i) > CP(c_j)$ then
8:        $p_j \leftarrow p_j + \triangle p$;
9:        Calculate $T_{max}$, i.e., the maximal interaction delay
          among all clients;
10:       if $T_{max} \leq \xi$ then
11:         $target \leftarrow c_j$, $flag \leftarrow 0$, $c(\triangle l) \leftarrow c_j$;
12:       else
13:         $p_j \leftarrow p_j - \triangle p$;
14:       end if;
15:     end if;
16:     if ($flag$) then
17:       $p_i \leftarrow p_i + \triangle p$;
18:       Calculate $\triangle T(c_i)$, i.e., the maximal interaction
          delay between $\triangle l$ and the other clients;
19:       if $\triangle T(c_i) \leq \xi$ then
20:         $target \leftarrow c_i$, $flag \leftarrow 0$, $c(\triangle l) \leftarrow c_i$;
21:       else
22:         $p_i \leftarrow p_i - \triangle p$;
23:       end if;
24:     end if;
25:   end for;
26:   if ($flag$) then
27:     $\triangle p \leftarrow \triangle p + 1$;
28:   end if;
29: end while;
30: return $c(\triangle l)$, $\triangle p$, $target$;
```

Based on Equations (5), (6), (7) and (8), Algorithm 2 gives the proxy assignment as well as the resource management approaches when a new client $\triangle l$ joins the system. Note that in *Step* 2 of Algorithm 2, we calculate the expected processing time as

$$r^*(c_i) = \begin{cases} \frac{a}{b^{p_i - q(c_i)}}, & p_i = q(c_i) \\ \frac{a}{b^{p_i - q(c_i) - 1}}, & p_i > q(c_i) \end{cases}. \quad (9)$$

The details of *Steps* 7 and 12 of Algorithm 2 are also given in Algorithm 3, in which CP denotes the cost performance of proxies. By expanding Equation (4) into a continuous function, we calculate CP as the differential coefficient of p_i, i.e.,

$$CP(c_i) = diff(\frac{r(p_i)}{p_i \cdot Cost(c_i)}, p_i) = \frac{-b^{q(c_i)-p_i}(1 + p_i \log b)}{{p_i}^2}. \quad (10)$$

It is worth mentioning that when $\triangle T_{min} > \xi$, the interaction latency of the newly joined client becomes the bottleneck. In this case, we need to add more cloud resources on the proxies. To this end, we add Algorithm 3 to increase the resources on the proxies. This cyclic iterative algorithm is designed to clarify: 1) which proxy should be used to handle the new client, 2) whether we should add more resources to this proxy, and 3) how many resources should be added.

On the other hand, Algorithm 4 is designed to handle the client departure, where $path^*$ denotes the longest path before client $\triangle l$ departs, T_{max} and T^*_{max} are the maximum interaction latency before and after client departure, respectively. We therefore summarize Theorem 2, of which the proof is detailed in Appendix D.

Algorithm 4 ClientDeparture()

```
1:  $q(c(\triangle l)) = q(c(\triangle l)) - 1$;
2:  $p(c(\triangle l)) = p(c(\triangle l)) - 1$;
3:  $path^* = LongestPath(G^*_A)$;
4:  Get $c_u$, $c_v$ from $path^*$;
5:  if $CP(c_u) \geq CP(c_v)$ then
6:    if $Cost(c_u) < Cost(c(\triangle l))$ then
7:      $p(c_u) = p(c_u) + 1$;
8:    end if;
9:  else
10:   if $Cost(c_v) < Cost(c(\triangle l))$ then
11:     $p(c_v) = p(c_v) + 1$;
12:   end if;
13: end if;
```

Note that according to Theorem 2, the client departure will not increase the maximum interaction latency. Hence, it is unnecessary to acquire proxy resources when considering the cost equality in *Steps* 6 and 10 of Algorithm 4.

B. Performance Validation

In this section, we aim to evaluate the performance of our resource provisioning design with user dynamic. As shown in Figure 17, we consider two typical patterns, *Stable* and *Flash-Crowd* [38] [39], in our evaluation. In particular, *Stable* means the number of clients is changing randomly between 30 to 50. *Flash-Crowd*, on the other hand, means the number of clients periodically increases from 5 to 50 and then decreases from 50 to 5 during a short period of time. These two patterns refer to typical "easy-to-handle" and "hard-to-handle" user demands, respectively.

Figure 18 shows the maximum interaction latency when user's delay tolerance is set to 450 ms (a typical RPG game). As we can see in the figure, our proposed online algorithm can successfully bound the latency under both *Stable* and *Flash-*

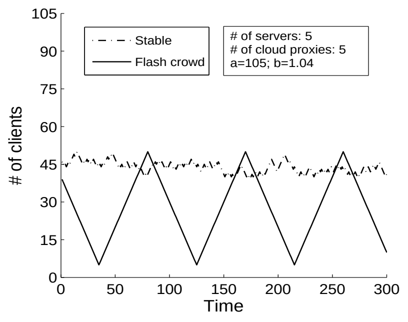


Fig. 17: Workload patterns

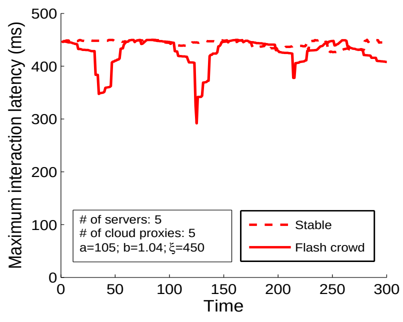


Fig. 18: Maximum interaction latency over time ($\xi = 450$ ms)

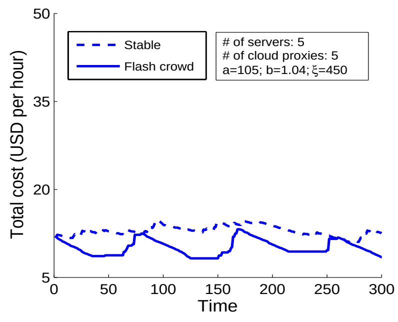


Fig. 19: Total cost over time ($\xi = 450$ ms)

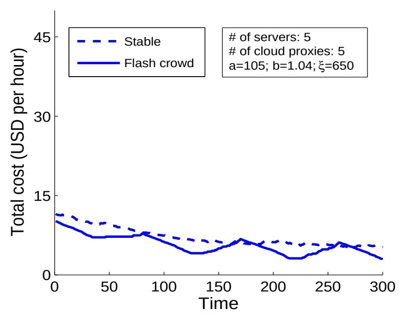


Fig. 20: Maximum interaction latency over time ($\xi = 650$ ms)

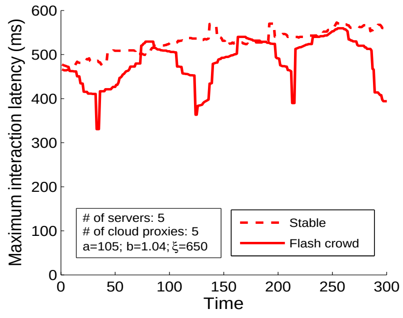


Fig. 21: Total cost over time ($\xi = 650$ ms)

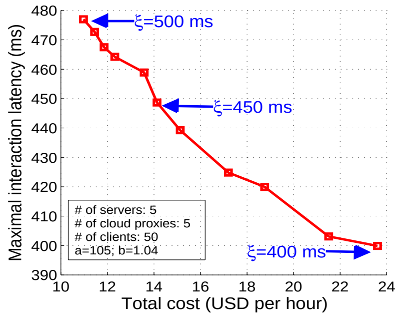


Fig. 22: Relationship between latency and cost

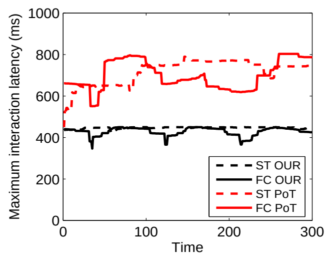


Fig. 23: Maximum interaction latency ($\xi = 450$)

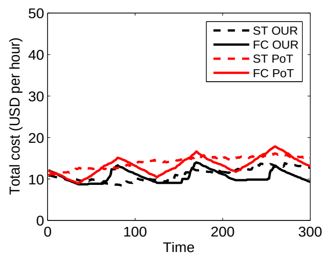


Fig. 24: Total cost ($\xi = 450$)

Crowd dynamics[13]. Figure 19 further presents the change of system cost. It is easy to see that the cost is less than 15 USD per hour (less than 3 cents/hour per user). Such a cost can be further reduced when we increase user's delay tolerance. In particular, Figure 21 shows that the latency can be further reduced to less than 5 USD per hour (less than 1 cent/hour per user) when users' delay tolerance is set to 650 ms. Although the interaction latency will be slightly increased, the maximum value is still bounded within 650 ms. As a summary, Figure 22 shows the relationship between cost and latency. With different configurations of ξ, our online protocol can always bound the maximum interaction latency with reasonable cost. It worth mentioning that since we dynamically manage resource when handling client arrival and departure, node state is updated in real time. Thus, when facing the fail-over (e.g., VM failure and migration), with the fast recovery capability of the cloud service provider, the proposed online protocol is capable to maintain the on-demand optimal allocations.

[13]The system's maximum interaction latency is highly related to the number of clients in the system. Such jitter may also affect the QoE on different clients. The related QoE optimization is, however, out of the scope of this paper.

To illustrate the effectiveness, we compare the performance of our design and a Power-of-Two [40] strategy. Power-of-Two is a well-known solution to Supermarket Model, where in a supermarket a customer joins the server with shorter queue from two randomly sampled servers. Power-of-Two is an efficiency strategy, which can achieve a near-optimal waiting time for the customers. Because of its simplicity and effectiveness, Power-of-Two has been extensively researched [41]–[43] and widely used [44], [45]. In this simulation, Power-of-Two serves as a baseline. When a new client arrives, Power-of-Two randomly samples two cloud proxies, and then assigns the one with the smaller maximum interaction latency to the client. Meanwhile, Power-of-Two increases resource on the selected proxy. Under both stable (ST) and flash crowd (FC) workload patterns when $\xi = 450$ ms, Figures 23 and 24 compare the maximum interaction latency and total cost of our solution (OUR) and Power-of-Two (PoT), respectively. As we can see, Power-of-Two induces larger maximum interaction latency as well as higher cost. Power-of-Two can hardly select the cloud proxy with minimum maximum interaction latency because of randomness. Besides, without considering the price/performance ratio when increasing resource, Power-of-Two tends to spend more money than our solution in operation.

VIII. Related Work

The origin of Distributed Interactive Applications (DIAs) can be traced back to 1983 when a United States research program initiated the SIMNET project [46] to train soldiers in battlefield tactics. Since then, an increasing number of academic, military and commercial DIA systems have been developed and documented. Nowadays, despite the increase of processing powers at participating clients and the availability of greater communication bandwidth, minimizing the interaction latency remains one of the most fundamental challenges in the DIA framework. Many studies have shown that the latency is particularly problematic when the network delays are comparable to the interaction time or speed [47]. Such studies suggested that the interaction latency should be bounded for real-world DIAs [48]. For example, the typical latency values to maintain real-time interaction fall between 40 and 300 ms [49]. Gutwin [33] investigated the latency effects on two types of user interactions: movement prediction and moving a shared object. This study showed that both gaming performance and user strategy will be greatly affected by interaction latencies higher than the expected range.

To minimize the interaction latency in DIAs, Webb *et al.* [50] proposed a nearest server assignment to reduce the client-server latency. Ta *et al.* [51] proposed a two-phase solution for large-scale DIAs. The study by Cronin *et al.* [31] further discussed the server placement problem to enhance user's interactivity. Vik *et al.* [32] explored the spanning tree problems in DIAs for latency reduction. Zhang *et al.* [12] revisited the problem and proposed a distributed-modify-assignment approach to adapt to the dynamics of client participation and network conditions. A recent study from Lee *et al.* [52] presents *Outatime*, a speculative execution system for mobile cloud gaming that is able to mask up to 120 ms of network latency.

For cloud computing, there have been a series of works measuring the performance of public or private cloud services from diverse aspects, including computation, storage, and networking services [53]–[56]. There are also many s-tudies addressing application designs that leverage cloud resources [57]–[63]. For example, Wu et al. [64] explored the use of cloud for Video-on-Demand applications [65], [66]. Huang *et al.* [67] provided an open-source cloud gaming system *GamingAnywhere*. Extensive experiments are performed to understand the video quality for both mobile and desktop clients. However, the deployment of enterprise cloud-based distributed interactive applications, which have emerged very recently, remains a mystery to the general public. Our study takes an initial step towards understanding this new generation of DIAs, in terms of its design, performance, and potential challenges.

IX. Conclusion and Further Discussion

In this paper, we examined the framework design and latency optimization in cloud-based distributed interactive applications through real system measurement and analysis. Our study identified the unique features as well as the fundamental design challenges in CDIA. However, there are still many open issues that can be further explored in this new-born system.

First, to better mitigate the interference between the computation-intensive and bandwidth-intensive tasks, we are working on the analysis of TCP/UDP flows on different types of VMs. Our preliminary result shows that VM *hypervisors* (also known as *virtual machine managers* such as Xen, KVM and VMware) and VM total capacities play important roles for the interference. On the other hand, except for the Amazon EC2, other representative cloud service providers such as GCE [26], IBM Cloud [68], Rackspace [69], Microsoft Azure [70], and AliCloud [27] have also played an important role in promoting the rapid development of CDIA. Our future work also includes the comparison between different cloud service providers regarding network performance.

Second, we are currently investigating the efficiency by directly migrating some DIA protocols/optimizations into the CDIA framework. This analysis helps to better enjoy the benefits of cloud computing while minimize the corresponding overheads. These investigations not only improve the overall performance of the CDIA framework, it also helps to better understand the development of many other cloud-based systems with similar design frameworks.

Third, handling real-world scaling issues is also an important part of our future work. The solution proposed in this paper requires global knowledge about the system state such as path latency and workload distribution. This may potentially affect the scalability of the solution. To mitigate this challenge, our on-going work is to build a hadoop2-like [71] [72] distributed information management models to collect global information, which enhances the efficiency and robustness of the system in a large-scale real-world environment.

Acknowledgment

This work was supported by National Natural Foundation of China (61472212), National Science and Technology Major Project of China (2015ZX03003004), National High Technology Research and Development Program of China (863 Program), EU Marie Curie Actions CROWN (FP7-PEOPLE-2013-IRSES-610524), NSERC Discovery Grant, E.W.R. Steacie Memorial Fellowship, and Industrial Canada Technology Demonstration Program (TDP) Grant. F. Wang's work was supported by a Start-up Grant from the University of Mississippi and H. Wang's work was supported by Chancellors Small Grant and Grant-in-aid programs from the University of Minnesota. Besides, grateful acknowledgement is made to Dr. Kun Yang and Liang Lv who gave me considerable help during experiments.

References

[1] Essential Facts about the Computer and Video Game Industry. http://www.theesa.com/facts/pdfs/ESA_EF_2012.pdf.
[2] Entertainment Business. http://www.entertainmentbusiness.nl/sites/default/files/documents/2012/Videogames.pdf.
[3] Gaikai. http://www.gaikai.com/.
[4] Onlive. http://www.onlive.com/.
[5] Ciinow. http://www.ciinow.com/.
[6] Gaikai in PlayStation4. http://www.theverge.com/2013/2/20/4010420/sonys-playstation-4-will-use-gaikai-game-streaming-technology.
[7] Cloud-based gaming planned for Microsoft Xbox 720. http://sonyps4playstation.com/cloud-based-gaming-planned-for-microsoft-xbox-720/.
[8] AMD Invests Into Ciinow. http://www.ubergizmo.com/2012/09/amd-invests-into-ciinow/.
[9] Jupyter. http://jupyter.org/.
[10] Shiny: A web application framework for R. http://shiny.rstudio.com/.
[11] F. Safaei, P. Boustead, C. Nguyen, J. Brun, and M. Dowlatshahi, "Latency-driven distribution: infrastructure needs of participatory entertainment applications," *IEEE Communications Magazine*, vol. 43, no. 5, pp. 106–112, 2005.
[12] L. Zhang and X. Tang, "Optimizing client assignment for enhancing interactivity in distributed interactive applications," *IEEE/ACM Transactions on Networking*, vol. 20, no. 6, pp. 1707–1720, 2012.

[13] M. Claypool and K. Claypool, "Latency and player actions in online games," *Communications of the ACM*, vol. 49, no. 11, pp. 40–45, 2006.
[14] Gaikai Powered Cloud-based Gaming on Samsung Smart TVs. http://www.engadget.com/2012/06/05/gaikai-powered-cloud-gaming-coming-to-samsung-smart-tvs/.
[15] S. A. Baset and H. G. Schulzrinne, "An analysis of the skype peer-to-peer internet telephony protocol," in *Proc. of IEEE INFOCOM*, 2006, pp. 1–11.
[16] Amazon EC2. http://aws.amazon.com/ec2/.
[17] Limelight Networks. http://www.limelight.com/.
[18] R. Dittner and D. Rule, *The Best Damn Server Virtualization Book Period*. Syngress, 2007.
[19] R. Jain and S. Routhier, "Packet trains–measurements and a new model for computer network traffic," *IEEE JSAC*, vol. 4, no. 6, pp. 986–995, 1986.
[20] R. Kawahara, E. K. Lua, M. Uchida, S. Kamei, and H. Yoshino, "On the quality of triangle inequality violation aware routing overlay architecture," in *Proc. of IEEE INFOCOM*, 2009, pp. 2761–2765.
[21] Networking and Traffic Control On Linux. http://tcng.sourceforge.net/.
[22] R. Shea, J. Liu, C. H. Ngai, and Y. Cui, "Cloud gaming: architecture and performance," *IEEE Network*, vol. 27, no. 4, pp. 16–21, 2013.
[23] K. T. Chen, Y. C. Chang, H. J. Hsu, D. Y. Chen, C. Y. Huang, and C. H. Hsu, "On the quality of service of cloud gaming systems," *IEEE Transactions on Multimedia*, vol. 16, no. 2, pp. 480–495, 2014.
[24] M. Claypool, D. Finkel, A. Grant, and M. Solano, "On the performance of onlive thin client games," *Multimedia Systems*, vol. 20, no. 5, pp. 471–484, 2014.
[25] M. Jarschel, D. Schlosser, S. Scheuring, and T. Hofeld, "An evaluation of qoe in cloud gaming based on subjective tests," in *Proc. of Innovative Mobile and Internet Services in Ubiquitous Computing*, 2011, pp. 330–335.
[26] GCE. https://cloud.google.com/compute/.
[27] AliCloud. https://www.aliyun.com/.
[28] Amazon web service. http://aws.amazon.com/.
[29] sysbench. https://github.com/akopytov/sysbench.
[30] iPerf. https://iperf.fr/.
[31] E. Cronin, S. Jamin, C. Jin, A. R. Kurc, D. Raz, and Y. Shavitt, "Constrained mirror placement on the internet," *IEEE JSAC*, vol. 20, no. 7, pp. 1369–1382, 2002.
[32] K.-H. Vik, P. Halvorsen, and C. Griwodz, "Multicast tree diameter for dynamic distributed interactive applications," in *Proc. of IEEE INFOCOM*, 2008, pp. 1597–1605.
[33] C. Gutwin, "The effects of network delays on group work in real-time groupware," in *Proc. of Springer ECSCW*, 2001, pp. 299–318.
[34] S. K. Barker and P. Shenoy, "Empirical evaluation of latency-sensitive application performance in the cloud," in *Proc. of ACM MMSys*, 2010, pp. 35–46.
[35] S. A. Ross, "Uses, abuses, and alternatives to the net-present-value rule," *JSTOR Financial management*, vol. 24, no. 3, pp. 96–102, 1995.
[36] C. Cotta and J. M. Troya, "A hybrid genetic algorithm for the 0–1 multiple knapsack problem," in *Springer Artificial neural nets and genetic algorithms*, 1998, pp. 250–254.
[37] H. Wang, R. Shea, X. Ma, F. Wang, and J. Liu, "On design and performance of cloud-based distributed interactive applications," in *Proc. of IEEE ICNP*, 2014, pp. 37–46.
[38] M. Mao and M. Humphrey, "Auto-scaling to minimize cost and meet application deadlines in cloud workflows," in *Proc. of ACM High Performance Computing, Networking, Storage and Analysis*, 2011, p. 49.
[39] K.-T. Chen, P. Huang, and C.-L. Lei, "Game traffic analysis: An mmorpg perspective," *Elsevier Computer Networks*, vol. 50, no. 16, pp. 3002–3023, 2006.
[40] M. Mitzenmacher, "The power of two choices in randomized load balancing," *IEEE TPDS*, vol. 12, no. 10, pp. 1094–1104, 2001.
[41] A. W. Richa, M. Mitzenmacher, and R. Sitaraman, "The power of two random choices: A survey of techniques and results," *Combinatorial Optimization*, vol. 9, pp. 255–304, 2001.
[42] A. Mukhopadhyay and R. R. Mazumdar, "Analysis of randomized join-the-shortest-queue (jsq) schemes in large heterogeneous processor-sharing systems," *IEEE Transactions on Control of Network Systems*, vol. 3, no. 2, pp. 116–126, 2016.
[43] A. Mukhopadhyay, A. Karthik, R. R. Mazumdar *et al.*, "Randomized assignment of jobs to servers in heterogeneous clusters of shared servers for low delay," *Stochastic Systems*, vol. 6, no. 1, pp. 90–131, 2016.
[44] X. Ren, G. Ananthanarayanan, A. Wierman, and M. Yu, "Hopper: Decentralized speculation-aware cluster scheduling at scale," *ACM SIGCOMM Computer Communication Review*, vol. 45, no. 4, pp. 379–392, 2015.
[45] S. Legtchenko, N. Chen, D. Cletheroe, A. I. Rowstron, H. Williams, and X. Zhao, "Xfabric: A reconfigurable in-rack network for rack-scale computers." in *Proc. of USENIX NSDI*, 2016, pp. 15–29.
[46] SIMNET. http://en.wikipedia.org/wiki/SIMNET/.
[47] P. M. Sharkey, M. D. Ryan, and D. J. Roberts, "A local perception filter for distributed virtual environments," in *Proc. of IEEE Virtual Reality Annual International Symposium*, 1998, pp. 242–249.
[48] M. Mauve, J. Vogel, V. Hilt, and W. Effelsberg, "Local-lag and timewarp: providing consistency for replicated continuous applications," *IEEE Transactions on Multimedia*, vol. 6, no. 1, pp. 47–57, 2004.
[49] C. Diot and L. Gautier, "A distributed architecture for multiplayer interactive applications on the internet," *IEEE network*, vol. 13, no. 4, pp. 6–15, 1999.
[50] S. D. Webb, S. Soh, and W. Lau, "Enhanced mirrored servers for network games," in *Proc. of ACM SIGCOMM NetGames*, 2007, pp. 117–122.
[51] D. N. B. Ta and S. Zhou, "A two-phase approach to interactivity enhancement for large-scale distributed virtual environments," *Elsevier Computer Networks*, vol. 51, no. 14, pp. 4131–4152, 2007.
[52] K. Lee, D. Chu, E. Cuervo, J. Kopf, Y. Degtyarev, S. Grizan, A. Wolman, and J. Flinn, "Outatime: Using speculation to enable low-latency continuous interaction for mobile cloud gaming," in *Proc. of ACM MobiSys*, 2015, pp. 151–165.
[53] S. Garfinkel, "An Evaluation of Amazon's Grid Computing Services : EC2 , S3 and SQS," *Harvard University Technical Report*, 2008.
[54] Y. Cui, Z. Lai, X. Wang, N. Dai, and C. Miao, "Quicksync: Improving synchronization efficiency for mobile cloud storage services," *IEEE TMC*, vol. PP, no. 99, pp. 1–1, 2015.
[55] Y. Cui, J. Song, K. Ren, M. Li, Z. Li, Q. Ren, and Y. Zhang, "Software defined cooperative offloading for mobile cloudlets," *IEEE/ACM Transactions on Networking*, vol. PP, no. 99, pp. 1–15, 2017.
[56] K. He, J. Chen, R. Du, Q. Wu, G. Xue, and X. Zhang, "Deypos: Deduplicatable dynamic proof of storage for multi-user environments," *IEEE Transactions on Computers*, vol. 65, no. 12, pp. 3631–3645, 2016.
[57] Y. Seung, T. Lam, L. E. Li, and T. Woo, "Cloudflex: Seamless scaling of enterprise applications into the cloud," in *Proc. of IEEE INFOCOM*, 2011, pp. 211–215.
[58] 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 Transactions on Circuits and Systems for Video Technology*, vol. 25, no. 12, pp. 1988–2001, 2015.
[59] S. Chen, Y. Shi, B. Hu, and M. Ai, "Mobility-driven networks (mdn): from evolution to visions of mobility management," *IEEE Network*, vol. 28, no. 4, pp. 66–73, 2014.
[60] T. Li, K. Wang, K. Xu, K. Yang, and H. Wang, "On efficient offloading control in cloud radio access network with mobile edge computing," in *Proc. of IEEE ICDCS*, 2017, pp. 2258–2263.
[61] T. Li, K. Xu, M. Sheng, H. Wang, K. Yang, and Y. Zhang, "Towards minimal tardiness of data-intensive applications in heterogeneous networks," in *Proc. of IEEE ICCCN*, 2016, pp. 1–9.
[62] K. Wang, K. Yang, H. H. Chen, and L. Zhang, "Computation diversity in emerging networking paradigms," *IEEE Wireless Communications*, vol. 24, no. 1, pp. 88–94, 2017.
[63] K. Wang, K. Yang, and C. Magurawalage, "Joint energy minimization and resource allocation in c-ran with mobile cloud," *IEEE Transactions on Cloud Computing*, vol. PP, no. 99, pp. 1–22, 2015.
[64] Y. Wu, C. Wu, B. Li, X. Qiu, and F. C. Lau, "Cloudmedia: When cloud on demand meets video on demand," in *Proc. of IEEE ICDCS*, 2011, pp. 268–277.
[65] K. Xu, T. Li, H. Wang, H. Li, Z. Wei, J. Liu, and S. Lin, "Modeling, analysis, and implementation of universal acceleration platform across online video sharing sites," *IEEE Transactions on Service Computing*, vol. PP, no. 99, pp. 1–15, 2016.
[66] K. He, J. Chen, Y. Zhang, R. Du, Y. Xiang, M. M. Hassan, and A. Alelaiwi, "Secure independent-update concise-expression access control for video on demand in cloud," *Information Sciences*, vol. 387, no. C, pp. 75–89, 2017.
[67] C.-Y. Huang, C.-H. Hsu, and K.-T. Chen, "Gaminganywhere: an open-source cloud gaming platform," *ACM SIG Multimedia Records*, vol. 7, no. 1, pp. 3–5, 2015.
[68] IBM Cloud. http://www.ibm.com/cloud-computing/.
[69] Rackspace. https://www.rackspace.com/.
[70] Microsoft Azure. https://azure.microsoft.com/.
[71] G. Turkington and G. Modena, *Learning Hadoop 2*. Packt, 2015.
[72] A. Sharma and S. Vyas, "Hadoop2 yarn," *IIJCS*, vol. 3, no. 9, 2015.
[73] I. Katriel, L. Michel, and P. Van Hentenryck, "Maintaining longest paths incrementally," *Springer Constraints*, vol. 10, no. 2, pp. 159–183, 2005.

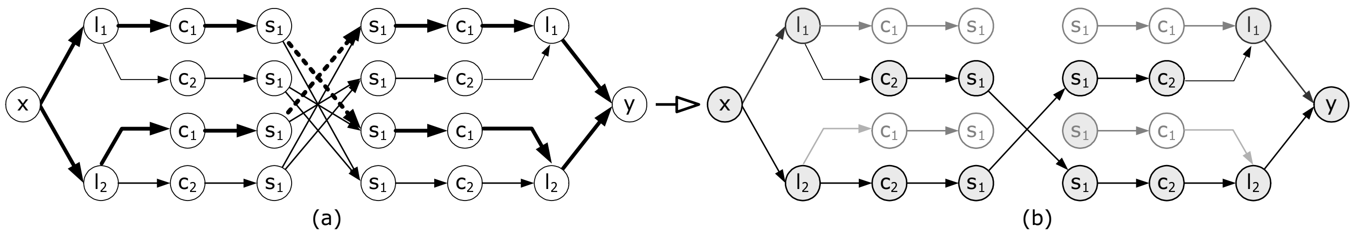


Fig. 25: Finding optimal assignment in converted DAG

Appendix A

To solve the assignment problem, we convert it into a directed acyclic graph (DAG) $G(V, E)$ with virtual source x and sink y (Figure 25). This is because the longest path (the slowest interaction path) can be found with worst-case running time of $O(|V| + |E|)$ [73] in a DAG $G(V, E)$. As illustrated in Figure 25(a) (which shows an example with 2 clients, 1 server and 2 cloud proxies), each path from x to y refers to one possible path between two clients in L. We first find the path with the highest latency. For example, in Figure 25(a), the longest paths are shown in the dark lines (there will be 2 longest paths in each round since they are symmetric). We then try to remove the edges between servers (dotted lines in Figure 25(a)) only when all client pairs are still connected after this removal. This step is repeated until no edge can be further removed from the graph. At last, we find A, the assignment of cloud proxy and server for each client in the remaining graph so that all the client pairs can be connected. The optimal algorithm is given in Algorithm 5. The proof of its optimality can be found in *Appendix B*.

Although the maximum interaction latency is bounded by the longest path, the optimal assignment A is not unique in Algorithm 5. This allows us to further improve other metrics, for example, the lease cost for cloud proxies. Note that the proposed optimal model assumes that the costs for all the cloud proxies are homogeneous. While this is partly valid for CDIAs that rely on their own cloud platform, e.g., Onlive, it is not the case for those using public clouds (e.g., Amazon EC2) with varying costs depending on such factors as location, time, and capacity. An extension of our model with heterogenous lease costs can be found in *Appendix C*.

Algorithm 5 OptimalLoadAssignment()

```
1: while $(IsConnected(L, G) == true)$ do
2:     $path^* = LongestPath(G)$;
3:     $Remove(path^*, G)$;
4: end while;
5: $Recover(path^*, G)$;
6: Return any viable $A$ from $G$;
```

Appendix B

Theorem 1. *Assignment A is an optimal assignment that minimizes the maximum interaction latency in Equation (3).*

Proof. Suppose that A' is another assignment in which the maximum interaction latency is smaller than A. Since A' and A can be both used to connect all client pairs in L, the longest path in A ($path^*$) can be replaced by a shorter path (say $path'$) that exists in A', and A can still make all clients in L connected after this replacement. Since $path'$ is shorter than the longest path in A, $path'$ also exists (not removed) in A's *residual graph* (the graph after the longest path removal at Algorithm 5 *Step* 3). This implies that there are two paths ($path^*$ and $path'$) in A's residual graph connecting identical client pairs with different latencies. This leads to a contradiction because the longest path in A can be safely removed without affecting the connectivity among clients.

Hence, the optimality of A is proved. □

Algorithm 6 AccommodateLeaseCost()

```
1: Sort $C$ by ascendant order of $Cost(c_i)$;
2: for all $i = 1$ to $n$ do
3:     $c_A(l_i) \leftarrow c_1$;
4:     $s_A(l_i) \leftarrow s_1$;
5: end for;
6: $i \leftarrow 1$;
7: while $(i <= n)$ do
8:     for all $j = 1$ to $i - 1$ do
9:         if $PathExisted(x, l_i, c_A(l_i), s_A(l_i),$
10:            $s_A(l_j), c_A(l_j), l_j, y, G) == false$ then
11:            break;
12:        end if;
13:    end for;
14:    if $j == i - 1$ then
15:        $i \leftarrow i + 1$;
16:    else
17:        if $Next(s_A(l_i), S)! = null$ then
18:            $s_A(l_i) \leftarrow Next(s_A(l_i), S)$;
19:        else if $Next(c_A(l_i), C)! = null$ then
20:            $c_A(l_i) \leftarrow Next(c_A(l_i), C)$;
21:            $s_A(l_i) \leftarrow s_1$;
22:        else
23:            $i \leftarrow i - 1$;
24:        end if;
25:    end if;
26: end while;
27: Return $A$;
```

Appendix C

For cloud proxies with heterogenous lease costs, we assume that the unit cost of a cloud proxy c_i (providing service to one client) is $Cost(c_i)$. For a given assignment A, we denote

$q_A(c_i)$ as the number of clients that are assigned to c_i. The overall lease cost $\mathbb{C}$ is therefore:

$$\mathbb{C} = \sum_{c_i \in C} Cost(c_i) \cdot q_A(c_i) \tag{11}$$

It is easy to see that the minimum cost can be found by searching all the assignments. This could be time-consuming in real practice. We therefore design a heuristic (Algorithm 6) to replace the last step in Algorithm 5 to obtain an economical assignment.

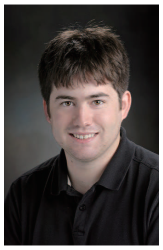
Appendix D

Theorem 2. *No matter whether $c(\triangle l) \in path^*$, releasing one unit of resource of $c(\triangle l)$ after client $\triangle l$ departing will not increase the maximum interaction latency, i.e., $T^*_{max} \leq T_{max}$.*

Proof. When $c(\triangle l) \notin path^*$, the value of $p(c(\triangle l)) - q(c(\triangle l))$ remains the same after releasing one unit of resource on $c(\triangle l)$ after client $\triangle l$ departing, according to Equation (4), the processing time on $c(\triangle l)$ remains the same. In this case, we have $T^*_{max} = T_{max}$. Otherwise when $c(\triangle l) \in path^*$, the longest path $path^*$ no longer exists after client $\triangle l$ departing, meanwhile the interaction latency between the other clients remains the same. In this case, we have $T^*_{max} < T_{max}$. In conclusion, no matter whether $c(\triangle l) \in path^*$, releasing one unit of resource on $c(\triangle l)$ after client $\triangle l$ departing will not increase the maximum interaction latency, i.e., $T^*_{max} \leq T_{max}$. □

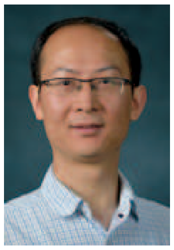
Haiyang Wang received the Ph.D. degree from Simon Fraser University, Burnaby (Metro-Vancouver), British Columbia, Canada, in 2013. He is currently an assistant professor in the Department of Computer Science at the University of Minnesota Duluth. His research interests include cloud computing, peer-to-peer networking, social networking, big data and multimedia communications.

Tong Li received the B.S. degree from the School of Computer Science of Wuhan University, Hubei, China in 2012. He received the Ph.D. degree from the Department of Computer Science and Technology of Tsinghua University, Beijing, China in 2017. He is now working as a research staff in Huawei Technologies. His research interests include network protocols, cloud/edge computing, network virtualization and resource management.



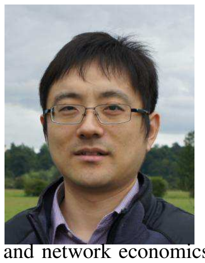
Ryan Shea received the Ph.D. degree in computer science from Simon Fraser University, Burnaby, BC, Canada, in 2016 and the certificate in university teaching and learning from Simon Fraser University, where he is currently working as a University Research Associate in Big Data Systems. His research interests include computer and network virtualization, and performance issues in cloud computing.

Xiaoqiang Ma is an assistant professor in the School of Electronic Information and Communications at Huazhong University of Science and Technology. He received B.S. degree from Huazhong University of Science and Technology, China, in 2010. He received M.S. and Ph.D. from Simon Fraser University, Canada, in 2012 and 2015, respectively. His current research interests are in the areas of wireless networking, multimedia, cloud, and big data.



Feng Wang Feng Wang (S'07-M'13) received both the B.S. degree and M.S. degree in Computer Science and Technology from Tsinghua University, Beijing, China in 2002 and 2005, respectively. He received the Ph.D. degree in Computing Science from Simon Fraser University, Burnaby, British Columbia, Canada in 2012. He is currently an Assistant Professor in the Department of Computer and Information Science at the University of Mississippi, University, MS, USA. He is a member of IEEE and a recipient of IEEE ICME Quality Reviewer Award (2011). He is a Technical Committee Member of Elsevier Computer Communications. He served as Program Vice Chair in International Conference on IOV 2014, and as TPC co-chair in IEEE CloudCom 2017. He also serves as TPC member in various international conferences such as IEEE INFOCOM, ICPP, IEEE/ACM IWQoS, ACM Multimedia, IEEE ICC, IEEE GLOBECOM and IEEE ICME.

Jiangchuan Liu (S'01-M'03-SM'08-F'17) is a University Professor in the Schoolof Computing Science, Simon Fraser University, British Columbia, Canada. He is an IEEE Fellow and an NSERC E.W.R. Steacie Memorial Fellow. He is an EMC-Endowed Visiting Chair Professor of Tsinghua University, Beijing, China and an Adjunct Professor of Tsinghua-Berkeley Shenzhen Institute. In the past he worked as an Assistant Professor at The Chinese University of Hong Kong and as a research fellow at Microsoft Research Asia. He received the B.S. degree (cum laude) from Tsinghua University, Beijing, China, in 1999, and the Ph.D. degree from The Hong Kong University of Science and Technology in 2003, both in computer science. He is a co-recipient of the inaugural Test of Time Paper Award of IEEE INFOCOM (2015), ACM SIGMM TOMCCAP Nicolas D. Georganas Best Paper Award (2013), and ACM Multimedia Best Paper Award (2012). His research interests include multimedia systems and networks, cloud computing, social networking, online gaming, big data computing, RFID, and Internet of things. He has served on the editorial boards of IEEE/ACM Transactions on Networking, IEEE Transactions on Big Data, IEEE Transactions on Multimedia, IEEE Communications Surveys and Tutorials, and IEEE Internet of Things Journal. He is a Steering Committee member of IEEE Transactions on Mobile Computing and Steering Committee Chair of IEEE/ACM IWQoS (2015-2017).



Ke Xu (M'02-SM'09) received his Ph.D. degree from the Department of Computer Science & Technology of Tsinghua University, where he serves as full professor. He is a co-inventor of three US patents and received the Best Paper Award of GLOBECOM 2015. He serves as Associate Editor of IEEE Internet of Things Journal and has guest edited several special issues in IEEE and Springer Journals. His research interests include next generation Internet, P2P systems, Internet of Things, network virtualization, and network economics. He is a member of ACM.