Premsankar, Gopika; Di Francesco, Mario; Taleb, Tarik

Edge Computing for the Internet of Things

Published in:
IEEE Internet of Things Journal

DOI:
10.1109/JIOT.2018.2805263

Published: 01/01/2018

Document Version
Peer reviewed version

Please cite the original version:
Premsankar, G., Di Francesco, M., & Taleb, T. (2018). Edge Computing for the Internet of Things: A Case Study. IEEE Internet of Things Journal, 5(2), 1275-1284. https://doi.org/10.1109/JIOT.2018.2805263

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JIOT.2018.2805263, IEEE Internet of Things Journal

Edge Computing for the Internet of Things: A Case Study

Gopika Premsankar, Mario Di Francesco, and Tarik Taleb

***Abstract*—The amount of data generated by sensors, actuators and other devices in the Internet of Things (IoT) has substantially increased in the last few years. IoT data are currently processed in the cloud, mostly through computing resources located in distant data centers. As a consequence, network bandwidth and communication latency become serious bottlenecks. This article advocates edge computing for emerging IoT applications that leverage sensor streams to augment interactive applications. First, we classify and survey current edge computing architectures and platforms, then describe key IoT application scenarios that benefit from edge computing. Second, we carry out an experimental evaluation of edge computing and its enabling technologies in a selected use case represented by mobile gaming. To this end, we consider a resource-intensive 3D application as a paradigmatic example and evaluate the response delay in different deployment scenarios. Our experimental results show that edge computing is necessary to meet the latency requirements of applications involving virtual and augmented reality. We conclude by discussing what can be achieved with current edge computing platforms and how emerging technologies will impact on the deployment of future IoT applications.**

***Keywords*—edge computing, fog computing, Internet of Things, mobile gaming.**

I. Introduction

There has been a substantial growth in the data generated by mobile and Internet of Things (IoT) devices. These devices (including sensors, smartphones and wearables) are characterized by limited computational and energy resources. Such limitations are currently addressed by offloading processing and storage from resource-constrained devices to the cloud [1]. Indeed, the cloud is an ideal solution for computation offloading due to its on-demand and scalable nature. However, cloud computing resources are hosted in large data centers built in locations far away from most end-users. This results in a high communication latency between end-users and the cloud. Moreover, the increasing amount of data exchanged adds substantial stress on the network links to the cloud.

To help address these issues, the concept of *edge* or *fog* computing has been proposed [2, 3]. According to this paradigm, computing resources are made available at the edge of the network, close to (or even co-located with) end-devices. Placing computing resources in close proximity to the devices generating the data reduces communication latency. Furthermore, network-intensive data can be processed and analyzed just one hop away from end-devices, thereby reducing the bandwidth demands on network links to distant data centers. The ease of processing and storing data close to the devices generating them will enable new services [4–7]. Finally, edge computing platforms support mobility of devices and geographically distributed applications [3]. Mobility and geographical distribution are indeed the key characteristics of IoT deployments that can particularly benefit from edge computing. A few representative applications include content delivery to vehicles, real-time analytics of data collected by mobile devices and environmental monitoring through geographically distributed wireless sensor networks.

The concept of bringing content closer to end-users is not new. Content delivery or distribution networks (CDNs) [8] deploy resources that replicate content from a source location onto servers close to the end-users. Information-centric networking (ICN) [9] is a similar approach for enhancing the Internet infrastructure to explicitly support content-based routing and forwarding. However, the CDN and ICN paradigms are limited to non-interactive content [10]; for instance, IoT data can be cached at the edge of the network [11, 12]. On the other hand, edge computing servers also provide computational capabilities and can host interactive applications that support user mobility. Furthermore, an edge computing platform can relieve privacy concerns as the data generated from IoT devices are stored and processed within nodes in the edge network. This means that data can be pre-processed to remove private information before being sent to the cloud [13]. Besides, offloading computation to resources closer to the users (and data centers that are not far away) can help reduce the energy consumption at the end-devices [14].

The contributions of this article are twofold. First, we classify and survey current edge computing architectures and platforms, then describe key IoT application scenarios that benefit from edge computing. Second, we carry out an experimental evaluation of edge computing and its enabling technologies in a selected use case represented by mobile gaming. Specifically, we demonstrate that edge computing is necessary to achieve satisfactory quality of experience in the considered use case. Indeed, mobile gaming relies on low latency and reliable communications as well as sensor data from mobile devices to create an immersive end-user

G. Premsankar and M. Di Francesco are with the Department of Computer Science, School of Science, Aalto University, Finland. E-mail: {gopika.premsankar, mario.di.francesco}@aalto.fi

T. Taleb is with Sejong University, Korea and with the School of Electrical Engineering, Aalto University, Finland. E-mail: tarik.taleb@aalto.fi

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JIOT.2018.2805263, IEEE Internet of Things Journal

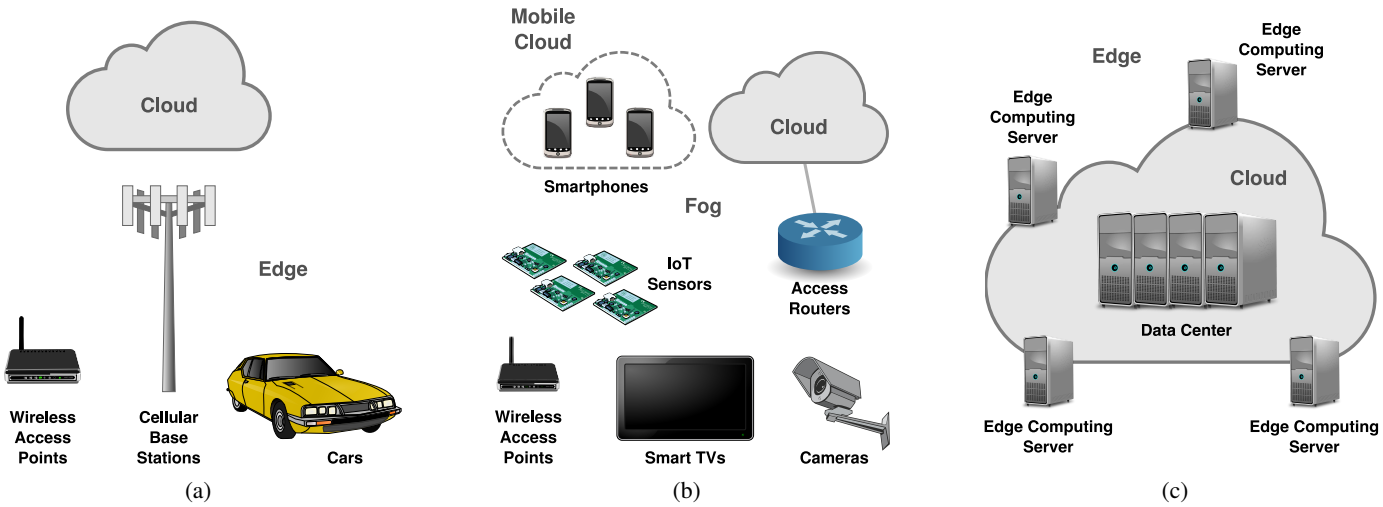

Fig. 1: Classification of edge computing platforms: (a) resource-rich servers deployed at the edge, (b) heterogeneous edge nodes and (c) edge-cloud federation. All devices not depicted within a cloud with a solid outline have edge computing capabilities.

experience [15, 16]. Pokémon Go[1] and Ingress[2] are examples of successful games that combine augmented reality and sensor information such as user location [17, 18]. Accordingly, we evaluate the benefits of edge computing for mobile gaming scenarios in this article. In particular, we consider gaming as a paradigmatic example of a larger class of applications that rely on rendering complex 3D environments, including virtual and augmented reality.

The rest of this article is organized as follows. Section II reviews and classifies existing edge computing platforms. Section III introduces the key enabling technologies behind edge computing. Section IV discusses the requirements of IoT applications and the benefits of edge computing for such applications. Section V presents the results of the performance evaluation of edge computing in mobile gaming. Section VI discusses our findings in relation to the currently available technologies. Finally, Section VII provides some concluding remarks and outlines possible directions for future research.

II. Edge Computing: Classes and Architectures

Different architectures have been proposed to realize edge computing platforms. A review of these architectures reveals that the edge of the network is not clearly defined and the nodes expected to participate at the edge can vary. Besides, the terminology used to describe the edge differs greatly, with the same term being used to define different architectures and functionality [22–24]. Thus, we begin by classifying the proposed architectures into three categories (Figure 1). The categorization is based on common features of deployments. However, in practice, features from one category can be used in combination with others. One category relies on *resource-rich servers deployed close to the end-devices*. Another group leverages resources from *heterogeneous nodes at the edge*, including the end-devices themselves. A third category is based on the *federation of resources at the edge and centralized data centers*. We detail the features of each category next; Table I summarizes our classification.

a) Resource-rich servers deployed at the edge: One option to realize an edge computing platform is to deploy resource-rich servers in the network to which end-users connect. Satyanarayanan et al. [2] present virtual machine (VM)-based *cloudlets* deployed on WiFi access points, one hop away from end-devices. A cloudlet is described as a "data center in a box" that offers a cluster of multicore computing capacity, storage and wireless LAN connectivity towards the edge. Ha et al. [14] propose a multi-tiered system using cloudlets to provide cognitive assistance for users. Video and sensor data collected from users through Google Glass are processed on the cloudlet to provide real-time assistance. Simoens et al. [25] present a scalable three-tier system using cloudlets for analytics and automated tagging of crowd-sourced video from user devices. Since the introduction of cloudlets, further research has proposed integrating cloudlets with femtocells, LTE base stations or even cars [13]. Greenberg et al. [26] describe the design of micro data centers consisting of thousands of servers and capable of hosting interactive applications for end-users. While these data centers have been used for deploying CDNs and email applications [26], they can be repurposed to host cloudlets [25]. Similar to cloudlets, Wang et al. [27] propose deploying a small set of servers on WiFi access points or a base station in the radio access network. The authors refer to this deployment as a micro cloud.

In the telecommunications ecosystem, *multi-access edge computing* (MEC) follows a similar approach of deploying

[1]http://www.pokemongo.com/
[2]http://www.ingress.com/

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JIOT.2018.2805263, IEEE Internet of Things Journal

Approach	Edge nodes	Edge network
Cloudlets [2]	Compact-size data centers deployed on WiFi access points, femtocells or LTE base stations	WiFi, 3G or LTE
Mobile cloudlets [13]	Compact-size data centers on cars	3G or LTE
Multi-access edge computing [19]	Servers deployed in the radio access network	3G, LTE, WiFi or other access technologies
Fog computing [3]	Heterogeneous nodes including high-end servers, routers, access points and set-top boxes	Multiple wireless access technologies including WiFi, 3G and LTE
Mobile cloud [20]	Neighboring mobile nodes form a cloud with one device chosen as resource coordinator	Local networking through WiFi or Bluetooth; Internet connectivity with WiFi, 3G and LTE
Edge cloud [21]	Compute or storage nodes deployed in the edge network and federated to cloud data centers	Home/enterprise networks and WiFi hotspots
FUSION [10]	Service nodes deployed on access points, local data centers and centralized data centers	Not defined

TABLE I: Summary of edge computing platforms and their features.

resource-rich resources at the edge. In this paradigm, cloud computing resources, storage and IT services are deployed in the radio access component of mobile networks. Such a platform consists of MEC servers integrated onto base stations or radio network controllers, while applications run on these servers through VMs. One of the key benefits of this architecture is the possibility to expose real-time radio link information to applications deployed at the edge. Since the initial white paper when MEC was known as mobile-edge computing [19], ETSI has expanded the scope of MEC to include access technologies other than mobile. A multi-access edge computing platform can also be deployed as a gateway in indoor environments and provide services including augmented reality, building management and social network applications [24].

b) Heterogeneous edge nodes: In contrast to the solutions described above, edge computing platforms can leverage a diverse set of computing resources. Bonomi et al. [3] propose a *fog platform* characterized by a highly virtualized system of heterogeneous nodes, ranging from resource-rich servers to more constrained edge routers, access points, set-top boxes and even end-devices (including smartphones and connected vehicles). The authors also recognize the heterogeneity of wireless connectivity as a key aspect of end-devices. Thus, different wireless access technologies are supported by their solution. A similar concept is presented in [28], with edge-devices (including smartphones and connected vehicles), routers and on-demand dedicated compute instances employed for processing data in the fog platform. Chiang and Zhang [29] describe a system which leverages computing resources on end-devices (including smartphones, Google Glass, home storage devices) and the cloud to carry out real-time data stream mining [30]. Nishio et al. [20] define fog computing as a cooperation-based *mobile cloud*, wherein heterogeneous mobile devices opportunistically share their resources to deliver services and applications. The proposed architecture consists of a local cloud formed by mobile devices in a neighboring area. These nodes can then share resources with other nodes in the same local cloud. One of the nodes is elected as a local resource coordinator and manages the allocation of tasks for services. These tasks can run on either devices in the local cloud or on the back-end cloud. Elkhatib et al. [31] propose the use of small and low-power computers such as Raspberry Pis to host fog services. The Raspberry Pis can be clustered together as independent and portable mini-clouds, which can be deployed in indoor or outdoor environments.

c) Edge-cloud federation: Another option to realize edge platforms is based on the federation of computing resources at the edge and centralized data centers. Chang et al. [21] describe this concept as an *edge cloud*. In their system, *edge apps* are used to deliver services at the edge as well as in distant cloud centers. The authors describe the use of edge apps to deploy indoor 3D localization and video monitoring applications. Similarly, Farris et al. [32] propose the federation of private and public clouds to enable integrated IoT applications. In this architecture, the edge node dynamically orchestrates the federation to maximize the number of executed tasks. Federation of clouds is also a key aspect of the FUSION architecture proposed by Griffin et al. [10]. In this paradigm, services are deployed on a cloud infrastructure distributed throughout the Internet. Application developers can deploy services in geographically distributed *execution zones* that can be located on IP routers, access points, base stations in the radio access network, and so on. Elias et al. [33] leverage an edge cloud (that mirrors public cloud services) along with a federated cloud to perform image classification with very low time and bandwidth requirements. The federated architecture and mirroring of the public cloud enables the use of existing open source repositories for machine learning and image classification at the edge.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JIOT.2018.2805263, IEEE Internet of Things Journal

III. Enabling technologies

The edge computing platforms described earlier are made possible by a few key enabling technologies, which are also crucial in the evolution of current mobile networks to their fifth generation (5G). In particular, 5G encompasses many new technologies addressing low latency and reliable communications, radio spectrum scarcity, energy-efficient operations and an increasing amount of data from heterogeneous devices. Besides, 5G networks are expected to support programmable and flexible deployments of services and core network functions through Network Function Virtualization (NFV) and Software-Defined Networking (SDN) [34]. These technologies are expected to play a key role in the development of edge computing platforms as well. A detailed survey of the enabling technologies is provided in [24]; we summarize the most important ones next and describe how they can be used at the edge.

a) Virtualization: Virtualization enables cloud computing providers to run multiple independent software instances on a single physical server. These instances can access the underlying physical resources while being isolated from each other. This isolation enables instances to run without interfering with (or even being aware of) other instances running on the same server. Currently, Virtual Machines (VMs) are the dominant means of deploying virtualized instances in cloud computing environments [35]. A software abstraction layer (hypervisor), lying between the VMs and the physical hardware, allows VMs to use the underlying CPU, storage and networking resources. Each VM runs its own guest operating system (OS) on top of the host server OS. Although hypervisor-based virtualization allows for excellent isolation of workloads and multi-tenancy, the hypervisor layer incurs in non-negligible overhead. Container-based virtualization has been proposed as a light-weight alternative to hypervisor-based virtualization. In this case, virtualized instances do not need to run a separate OS and can share the resources of the underlying host OS. Modifications are made to the OS to ensure isolation between containers. This form of virtualization allows reducing instance start times and generally results in better performance [34].

Apart from sharing physical resources, virtualization also allows the migration of VMs or containers. More specifically, migration consists in moving computing resources from one physical server to another. This is very useful for several scenarios, including the consolidation of virtualized instances to reduce data center energy consumption or to adapt to user mobility. *Live migration* is a technique that reduces the time during which a virtual instance is not accessible as it is being moved from one server to another. While live migration of VMs has existed for a long time and is being used extensively by cloud providers, migration of containers is still relatively new. Virtualization and live migration techniques are particularly important for edge computing platforms. These virtualization technologies are also the basis of NFV and SDN as described next.

b) Network Function Virtualization (NFV) and Software-Defined Networking (SDN): NFV [36] involves the implementation of network functions as software modules that can run on general-purpose hardware. It decouples the software from the underlying hardware by leveraging the previously described virtualization technologies. With this approach, different network functions and services no longer need to run on dedicated hardware. Instead, they can be executed on general-purpose nodes. NFV also offers improved flexibility, as virtualized network functions can be deployed in the location most suited for efficient delivery of mobile applications and services.

SDN [37] complements NFV by decoupling the management or control plane from the data plane over which data packets are forwarded. SDN enables easier and more flexible management of networks through abstractions and a logically centralized controller that handles policy and forwarding decisions. Moreover, the softwarization of the controller allows for faster deployment of new services. SDN, together with NFV, enables flexible and programmable deployment of software-based modules, thereby simplifying network configuration and management. Besides, these technologies are extremely important for network operators to quickly deploy new software functions with a limited cost. For instance, NFV enables automated deployment of virtual resources to meet a sudden increase in the traffic generated by an IoT application at a certain location. Jointly with the availability of edge computing platforms, NFV can bring the needed virtual resources close to end-users, for instance, on equipment within their premises. On top of this, SDN enables automated orchestration of virtualized instances as well as flexible policy control and routing of the increased traffic to the newly deployed resources.

c) Computation offloading: Typically, computation and storage are offloaded from resource-constrained mobile devices to the cloud [1], i.e., processing-heavy tasks are sent for execution in the cloud which, in turn, sends the results back to the devices [38]. Clearly, offloading can involve edge computing platforms instead of (or in addition to) the centralized cloud. In either case, end-devices access cloud resources as a thin client or through a web browser [39]. Offloading computation from devices has several benefits. For instance, the battery life of resource constrained end-devices can be extended by avoiding complex local processing. Offloading to the edge instead of the cloud results in even lower energy consumption at the end-device [24, 25]. Moreover, offloading computation enables several types of applications to run on resource-constrained devices, including mobile gaming, mobile learning, natural language processing and mobile healthcare.

IV. Edge computing for IoT Applications

The IoT is characterized by resource-constrained devices such as sensors, smartphones, wearable devices and machines connected to the Internet. The IoT lays out a foundation for the digitalization of the physical world that can be described in terms of machine-friendly data. Once sampled or generated, these data can be automatically processed and interpreted to provide innovative services in diverse areas ranging from mobile healthcare, to smart power generation and intelligent transportation systems. Several distinctive features of the IoT make it well suited for deployments based on edge computing platforms [3]. The characteristics and some of the representative use cases enabled by edge computing are the following.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JIOT.2018.2805263, IEEE Internet of Things Journal

- **Low-latency communication** is critical for several IoT applications, including connected vehicles, mobile gaming, remote health monitoring, warehouse logistics and industrial control systems. These scenarios are characterized by real-time actions or responses based on processing data generated by end-devices.
- An increasing amount of data generated by IoT deployments today are **bandwidth-intensive**, including video from surveillance cameras, police patrol cars and user devices. Placing computational resources one-hop away from high-bandwidth data sources implies that less data need to be sent to the distant cloud data centers [13]. For instance, videos and sensor data from hazardous locations can be processed locally to provide real-time information to responders in public safety applications [40].
- **Geographical distribution** is a key characteristic of IoT applications based on sensor networks. Indeed, the related use cases highly benefit from processing data locally through edge computing platforms. One example is given by collision avoidance systems deployed at the edge of vehicular networks, e.g., at the roadside units deployed for communication purposes. These systems depend on sensor data (such as location, velocity, acceleration, and so on) generated by both vehicles and pedestrians. Processing the data locally replaces sending data to distant cloud data centers, thus achieving low-latency communications.
- **Device mobility** places additional demands on low-latency processing of device data. Indeed, edge computing platforms support migration of virtualized resources based on the mobility of end-devices, thus allowing the data generated by these devices to be processed locally and with a satisfactory quality of experience.

The current state of the art in the IoT largely involves device-driven (i.e., machine-type) communications, and does not require explicit human intervention. The next step forward is the tight integration of digital data and physical environments through real-time wireless communications. An emerging class of applications involve the integration of sensor- or user-generated inputs and artificially-created 3D scenarios. For instance, samples generated by sensor nodes may be used to replicate the movements and gestures of a person in a specific physical setting onto a virtual environment.

The real-time nature of interactive applications is deeply connected with perception and actuation. The response time of human beings for different types of stimuli varies depending on the specific circumstances. However, it can be as low as a few milliseconds for situations where rapid actions are taken or expected. The response time of an artificial system significantly affects the user experience of human-computer interactions too. For instance, there are very stringent latency and reliability constraints for immersive applications, such as virtual and augmented reality [41, 42]. Indeed, virtual reality builds an immersive environment that adapts to the movements and the interactions of users. Augmented reality extracts information about the physical world and overlays it on the field of view of a person, i.e., on a head-mounted display. The recent success of Pokémon Go highlights the popularity of games that incorporate the physical location of the user and augmented reality into mobile games. Future games are expected to combine additional wearable sensor data to enable context-aware and more immersive gaming experiences. The response times of such games significantly affect the user experience [43, 44]. The response time becomes especially important in multi-user games whose gameplay can be significantly affected by delays in processing user- or sensor-generated input.

In the cases outlined above, the application needs to create a virtual environment and adapt it according to the user inputs. Such an environment needs to be realistic, thereby involving complex 3D modeling and rendering operations that are resource-intensive. The cloud computing paradigm offers abundant computation and storage capabilities that can be leveraged to realize close to photorealistic rendering in a very short time. However, cloud resources can be located far away from the user, thus introducing significant delays. While they would be acceptable for non-interactive applications, these delays cannot be tolerated in the considered use cases, as they would result in a poor user experience.

The edge computing paradigm offers the same key features of the cloud in a location close to the user, thus resulting in a much shorter latency. An important question is: *what can edge computing achieve with the current state of the art*? To address this question, we focus on the specific use case of mobile gaming next.

V. Use Case: Mobile Gaming

We carry out an experimental evaluation of mobile gaming using a prototype edge computing platform. This use case is particularly meaningful given the currently available technologies. For our evaluation, we select Neverball[3], an open source 3D arcade game wherein the player controls a ball by tilting the floor so as to collect coins and reach a designated exit point through several levels. Neverball is representative of a larger class of applications that rely on rendering complex 3D environments, including virtual and augmented reality. Moreover, due its peculiar gameplay, it can be used as a fitness game, for instance, by employing balance boards (including Nintendo's Wii Fit) as input devices [45]. In such scenarios, the end-to-end latency is extremely important as the user needs very quick responses from the server to have a good quality of experience. In our scenario, we use a mobile phone (end-device) that sends the relevant game input to the gaming server, which in turn renders the content and streams the video back to the end-device.

A. Testbed Setup

In our evaluation, we carry out experiments by using the open-source GamingAnywhere cloud gaming platform [46]. We focus on the *response delay* and use the same methodology described by Huang et al. [46]. Specifically, we define the response delay as the time elapsed between an action performed by the user and the occurrence of the corresponding outcome

[3] http://neverball.org

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JIOT.2018.2805263, IEEE Internet of Things Journal

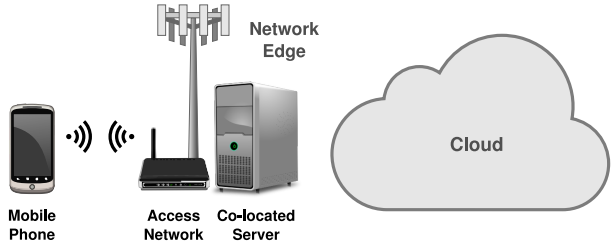

Fig. 2: Testbed setup used for the network edge scenario.

at the client device. The response delay includes three main components:

- processing delay (PD): the time taken by the server to process the user input and render the corresponding frame;
- playout delay (OD): the time taken by the client to decode and display the frame on its own screen;
- network delay (ND): the round trip time (RTT) between the client and the server, which accounts for the time taken to send data from the client to the server and back.

In our experiments, we consider two access technologies, i.e., WiFi and LTE. In both cases, we use our university network, in particular, the NetLeap 4G network provided by Nokia Solutions and Networks for LTE. The client device is a Google Nexus 5 mobile phone running Android 5.1.1. The gaming server is deployed on a workstation with a 4-core Intel Xeon E3-1230 CPU, 16 GB of RAM and two NVIDIA Quadro 2000 GPUs. We consider three different server deployment scenarios: a local deployment at the network edge, a special-purpose cloud computing infrastructure, and a commercial public cloud provider.

Specifically, the network edge scenario is represented by a server co-located with an LTE base station and a workstation deployed in the same wireless network as the client device for WiFi, as illustrated in Figure 2. This setup corresponds to the architecture presented in Figure 1a, wherein computing resources are deployed in the edge network. The second considered scenario, instead, consists of a special-purpose cloud infrastructure, namely, the cPouta service offered by CSC (the Finnish IT center for science), running OpenStack and located in Kajaani, Finland. Finally, the last scenario uses a commercial public cloud provider, i.e., Amazon; in this case, we consider the two geographically closest EC2 data center locations, i.e., Frankfurt and Ireland.

The Neverball game runs on the Ubuntu 14.04.4 Linux operating system under three different configurations: bare metal, i.e., directly on the host OS without any virtualization technology; within a Linux container; within a virtualized instance. We employ Docker (version 1.10.3) to run a container and QEMU (version 2.5.0) to run a VM on the host. In each case, we assign one of the GPUs of the host to the container or VM. For the server deployment on Amazon EC2, we use two types of GPU instances, both featuring Intel Xeon E5-2670 processors and NVIDIA GRID K520 GPUs: g2.2xlarge (with 1 GPU and 8 vCPUs) and g2.8xlarge (with 4 GPUs and 32 vCPUs). We use dedicated instances, running on reserved physical servers and isolated from others, to reduce the impact of the data-center load on the performance.

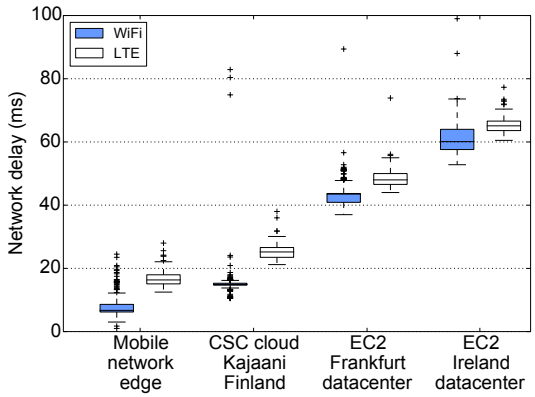

Fig. 3: Impact of network access technologies and server deployment on the network delay.

We ran four iterations of each experiment to characterize the statistical significance of the obtained results. We replayed gaming sessions that were previously recorded to achieve consistent outcomes that are not considerably affected by the timing of the input events. Individual sessions lasted from 1 to 3 minutes. We configured GamingAnywhere to stream videos at 30 FPS with a bitrate of 4.5 Mbps for all experiments.

B. Experimental Results

We first study the impact of server deployment on the network delay. Then, we examine the overhead of different virtualization technologies and how the screen resolution affects the response delay. Finally, we quantify the impact of additional computational resources offered by the cloud on the processing delay.

Figure 3 shows the network delay for the considered access technologies (i.e., WiFi and LTE) as a function of the location of the server deployment as a box plot. To obtain these results, we measured the network delay in terms of the RTT between the client and the server obtained through ICMP ping messages. The ping measurements were carried out at regular intervals over a duration of at least 6 hours.

The figure clearly shows the impact of the data center on the network delay: the two scenarios with servers located in Finland obtain significantly lower values (i.e., consistently below 25 ms) than those of the data centers located in other countries, with a geographical distance[4] of about 1,500-2,000 km. The edge network scenario allows a network delay of less than 20 ms over LTE, which can be considered the state of the art of currently available wireless communication technologies. A public cloud incurs in a delay that is at least twice as much, i.e., 50 ms in the best case. To a certain extent, the network delay of the CSC cloud is quite low, but this is also related by the high-speed network connecting our university with the data-center

[4]In contrast, Kajaani is less than 500 km away from Helsinki.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JIOT.2018.2805263, IEEE Internet of Things Journal

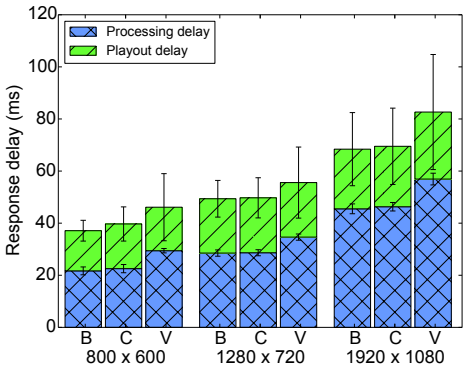

Fig. 4: Impact of the resolution on the response delay for different configurations: bare metal (B), containers (C) and virtual machines (V).

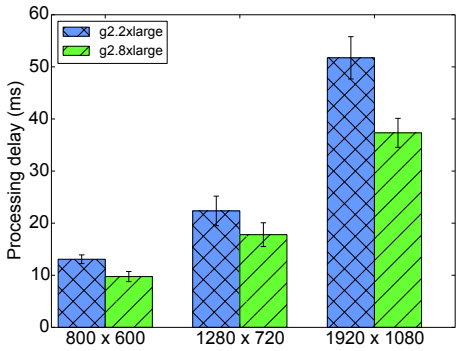

Fig. 5: Average processing delay as a function of the resolution for different Amazon GPU instances.

in Kajaani – similar results would be difficult to achieve with standard cloud providers. Besides, WiFi obtains shorter delays than LTE on the average, even though the difference becomes small when the data center is far. Moreover, the variance of the results obtained with WiFi is significantly higher than that with LTE. This implies that the jitter is lower over LTE, making it actually a better choice for our use case, as it involves video streaming.

Figure 4 shows the components of the response delay that are not affected by the network (i.e., the processing and the playout delay) for different virtualization technologies as a function of the screen resolution. Specifically, the bare metal, container, and VM configurations are indicated as B, C, and V (respectively) in the figure. The figure demonstrates that the performance of containers is almost the same as the bare metal configuration, irrespective of the video resolution. Hypervisor-based virtualization, instead, incurs in about 30% more processing delay. This additional delay is significant for the considered scenario: for instance, a resolution of 1280×720 pixels can be actually streamed at the target value of 30 FPS with containers, but not with VMs. With the highest resolution of 1920×1080 pixels, none of the considered configurations allows to match the desired frame rate.

The figure also shows that the playout delay at the client is significant: it is actually comparable to the processing delay when the resolution is not full HD. The playout delay increases with the screen resolution too, even though it never exceeds 25 ms on the average. The variance of the playout delay is much higher than that of the processing delay, increases with the resolution and is higher when VMs are used. This can be related to the higher variance in the content of the source video content before compression, and on the additional jitter affecting communications when using hypervisor-based virtualization.

Next, we evaluate whether the use of more powerful computing resources in the cloud (thereby reducing the processing delay) can compensate for the network delay in accessing these resources. Figure 5 shows the average processing delay on two different types of EC2 instances as a function of the screen resolution (we do not present the playout delay as it is similar to the results obtained earlier). The g2.2xlarge instance achieves shorter processing delays than those for the bare metal configuration for the resolutions of 800×600 pixels and 1280×720 pixels. However, at the highest resolution of 1920×1080 pixels, the g2.2xlarge instance takes on average 5 ms longer than the bare metal workstation. With the more powerful g2.8xlarge instance, the processing delay is shorter than the bare metal configuration for all resolutions. These results indicate that the additional computational resources offered by the cloud are not effective for the full HD resolution. In fact, most of the processing delay is due to the encoding of the source video content, as opposed to rendering. At 800×600 pixels, the g2.2xlarge is on average 12 ms faster than the bare metal workstation and at 1920×1080 pixels, 8 ms faster. Even though the reduction in the processing delay is significant as a percentage (i.e., both GPU instances are twice as fast as the bare metal configuration), the actual gain is rather limited in terms of the sheer numbers.

VI. Discussion

Although we have focused on the use case of gaming, the results of our experimental evaluation offer insightful considerations on using edge computing for next-generation IoT applications. It is clear that hosting computing resources very close to the end-users, possibly at the access network edge, is the only viable option to achieve a satisfactory quality of experience. While a response delay below 150 ms is generally considered acceptable for interactive applications, fast-paced interactions cannot tolerate delays beyond 70 ms [43]. By combining the components of the response delay shown in Figures 3 and 4, it is apparent how fast-paced interactions cannot be satisfactory when using any of the data centers in the considered public cloud. Indeed, the mobile network edge configuration allows to play the game at a HD resolution with processing times below 70 ms; not even using the CSC cloud allows this result. As demonstrated by Figure 5, the availability of much higher computational resources in the cloud (rather

than at the edge) does not help, as the gain in processing delay is not enough to overcome the additional network delay.

Our experiments involved only one end-user. However, the key observations from our experimental evaluation remain the same even for scenarios involving multiple users. In fact, the considered use cases relies on one VM (or container) per user, as it usually happens in mobile (cloud) gaming [46]. In the case of online multi-player gaming, a single virtualized instance (container or VM) is deployed for rendering per user at the edge [47]. Thus, the fundamental observations from our experiments still hold. Moreover, traffic flow control within the data center networks (at the edge) can be used to reduce the service latency between different components for a large number of users [48]. Indeed, issues related to network access would become more critical in a scenario involving a large number of users. In this case, edge computing platforms can leverage virtualization, NFV and SDN to scale out resources when the number of end-users increases. Specifically, NFV can be employed to deploy virtualized gaming modules at the edge and exploit real-time information on the access network to appropriately tune application parameters (for instance, video encoding parameters in the cloud gaming use case). Edge orchestrators can also launch VMs (or containers) with higher processing capabilities based on the requirements of the specific games. Finally, SDN can be employed to manage networking at the edge and flexibly control network flows. Furthermore, mobility of users can be handled through live migration of edge computing resources (VMs or containers).

Our evaluation has considered rather limited options concerning the availability and location of devices with suitable computing capabilities. However, the number of these devices is rapidly increasing in the infrastructure of mobile network operators; a similar trend is expected for content and service providers as well. As a result, selection and management of edge nodes will become a crucial aspect for successful deployment of next-generation IoT applications. That is exactly where NFV and SDN come into play [49]. Complex real-time resource allocation and optimization problems will emerge, beyond the current state of the art [50].

The delays incurred by current technologies are the main limiting factor of the performance that can be achieved today. Indeed, it is impossible to lower the response time to 10 ms or even less without fundamental advances in both wireless communication and computing technologies. While one of the objectives in the 5G initiative is to reduce latency to one millisecond, it is still unclear whether a significant reduction of the factors affecting delay beyond the access network can be achieved without a fundamental breakthrough. What we can certainly confirm is that edge computing is definitely needed to move forward in achieving such ambitious goals.

VII. Conclusion

This article has investigated the suitability of edge computing for emerging IoT applications. Specifically, we evaluated the performance of edge computing for mobile gaming as a representative scenario of new applications incorporating physical sensory inputs in addition to those explicitly generated by the user. The obtained results have shown that edge computing is necessary to enable fast-paced interactive games. Although regional data centers allow to significantly reduce network latency, only hosting resources at the edge enables a satisfactory quality of experience for gaming. Furthermore, increasing the computational capabilities of the servers in the cloud does not compensate for the increase in network latency. Therefore, deploying even limited computing resources at the edge helps improve the quality of experience in the considered use case. An interesting future research direction is represented by a large-scale evaluation of mobile gaming based on the edge computing paradigm. Such a study could consider, for instance, online multi-player games. It would also be interesting to compare the performance of different edge computing architectures for specific application scenarios. We hope that this article will encourage further research in this direction.

Acknowledgments

The authors would like to thank Teemu Käämäräinen, Jaakko Kotimäki, Janne Savikko and Lauri Tirkkonen for their help in the setup of the experiments. This work was partially supported by the Academy of Finland grants number 299222 and number 305507.

References

[1] H. T. Dinh, C. Lee, D. Niyato, and P. Wang, "A survey of mobile cloud computing: architecture, applications, and approaches," *Wirel. Commun. Mob. Comput.*, vol. 13, no. 18, pp. 1587–1611, Dec 2013.

[2] M. Satyanarayanan, P. Bahl, R. Caceres, and N. Davies, "The case for VM-based cloudlets in mobile computing," *IEEE Pervasive Computing*, vol. 8, no. 4, pp. 14–23, Oct.-Dec. 2009.

[3] F. Bonomi, R. Milito, P. Natarajan, and J. Zhu, "Fog computing: A platform for Internet of Things and analytics," in *Big Data and Internet of Things: A Roadmap for Smart Environments*. Springer, Mar 2014, pp. 169–186.

[4] R. Atat, L. Liu, H. Chen, J. Wu, H. Li, and Y. Yi, "Enabling cyber-physical communication in 5g cellular networks: challenges, spatial spectrum sensing, and cyber-security," *IET Cyber-Physical Systems: Theory & Applications*, vol. 2, no. 1, pp. 49–54, 2017.

[5] K. Wang, Y. Wang, Y. Sun, S. Guo, and J. Wu, "Green industrial internet of things architecture: An energy-efficient perspective," *IEEE Communications Magazine*, vol. 54, no. 12, pp. 48–54, 2016.

[6] J. Wu, S. Guo, J. Li, and D. Zeng, "Big data meet green challenges: Big data toward green applications," *IEEE Systems Journal*, vol. 10, no. 3, pp. 888–900, 2016.

[7] ——, "Big data meet green challenges: Greening big data," *IEEE Systems Journal*, vol. 10, no. 3, pp. 873–887, 2016.

[8] P. Frangoudis, L. Yala, A. Ksentini, and T. Taleb, "An architecture for on-demand service deployment over a telco CDN," in *Proc. IEEE ICC '16*, Kuala Lampur, Malaysia, May 2016.

[9] B. Ahlgren, C. Dannewitz, C. Imbrenda, D. Kutscher, and B. Ohlman, "A survey of information-centric networking," *IEEE Communications Magazine*, vol. 50, no. 7, pp. 26–36, 2012.

[10] D. Griffin, M. Rio, P. Simoens, P. Smet, F. Vandeputte, L. Vermoesen, D. Bursztynowski, and F. Schamel, "Service oriented networking," in *EuCNC 2014*, June 2014.

[11] R. Li, H. Harai, and H. Asaeda, "An aggregatable name-based routing for energy-efficient data sharing in big data era," *IEEE Access*, vol. 3, pp. 955–966, 2015.

[12] R. Li, H. Asaeda, and J. Li, "A distributed publisher-driven secure data sharing scheme for information-centric iot," *IEEE Internet of Things Journal*, 2017.

[13] M. Satyanarayanan, P. Simoens, Y. Xiao, P. Pillai, Z. Chen, K. Ha, W. Hu, and B. Amos, "Edge analytics in the Internet of Things," *IEEE Pervasive Computing*, vol. 14, no. 2, pp. 24–31, Apr.-June 2015.

[14] K. Ha, Z. Chen, W. Hu, W. Richter, P. Pillai, and M. Satyanarayanan, "Towards wearable cognitive assistance," in *Proc. 12th Int'l. Conf. Mobile Systems, Applications, and Services*, Bretton Woods, NH, June 2014.

[15] S. Liszio and M. Masuch, "Designing shared virtual reality gaming experiences in local multi-platform games," in *International Conference on Entertainment Computing*. Springer, 2016, pp. 235–240.

[16] J. Farman, "Locative life: Geocaching, mobile gaming, and embodiment," *Digital Arts and Culture 2009*, 2009.

[17] S. Chess, "Augmented regionalism: Ingress as geomediated gaming narrative," *Information, Communication & Society*, vol. 17, no. 9, pp. 1105–1117, 2014.

[18] E. Graells-Garrido, L. Ferres, and L. Bravo, "The effect of Pokémon Go on the pulse of the city: A natural experiment," *arXiv preprint arXiv:1610.08098*, 2016.

[19] ETSI, "Mobile-Edge Computing Introductory Technical White Paper," Sep. 2014.

[20] T. Nishio, R. Shinkuma, T. Takahashi, and N. B. Mandayam, "Service-oriented heterogeneous resource sharing for optimizing service latency in mobile cloud," in *ACM Proc. 1st Int'l. Workshop Mobile Cloud Computing & Networking (MobileCloud '13)*, Bangalore, India, July 2013.

[21] H. Chang, A. Hari, S. Mukherjee, and T. Lakshman, "Bringing the cloud to the edge," in *IEEE Conf. Comp. Comm. Workshops (INFOCOM WKSHPS)*, April 2014.

[22] P. Mach and Z. Becvar, "Mobile edge computing: A survey on architecture and computation offloading," *IEEE Communications Surveys & Tutorials*, 2017.

[23] Y. Mao, C. You, J. Zhang, K. Huang, and K. B. Letaief, "A survey on mobile edge computing: The communication perspective," *IEEE Communications Surveys & Tutorials*, 2017.

[24] T. Taleb, K. Samdanis, B. Mada, H. Flinck, S. Dutta, and D. Sabella, "On multi-access edge computing: A survey of the emerging 5g network edge architecture & orchestration," *IEEE Communications Surveys & Tutorials*, 2017.

[25] P. Simoens, Y. Xiao, P. Pillai, Z. Chen, K. Ha, and M. Satyanarayanan, "Scalable crowd-sourcing of video from mobile devices," in *Proc. of the 11th annual International conference on Mobile systems, applications, and services*. ACM, 2013, pp. 139–152.

[26] A. Greenberg, J. Hamilton, D. A. Maltz, and P. Patel, "The cost of a cloud: research problems in data center networks," *ACM SIGCOMM computer communication review*, vol. 39, no. 1, pp. 68–73, 2008.

[27] S. Wang, R. Urgaonkar, T. He, M. Zafer, K. L. Chan, and K. K. Leung, "Mobility-induced service migration in mobile micro-clouds," in *Military Communications Conference (MILCOM), 2014 IEEE*. IEEE, 2014, pp. 835–840.

[28] K. Hong, D. Lillethun, U. Ramachandran, B. Ottenwälder, and B. Koldehofe, "Mobile fog: A programming model for large-scale applications on the internet of things," in *Proceedings of the second ACM SIGCOMM workshop on Mobile cloud computing*. ACM, 2013, pp. 15–20.

[29] M. Chiang and T. Zhang, "Fog and IoT: An overview of research opportunities," *IEEE Internet of Things Journal*, vol. PP, no. 99, pp. 1–1, 2016.

[30] L. Canzian and M. Van Der Schaar, "Real-time stream mining: online knowledge extraction using classifier networks," *IEEE Network*, vol. 29, no. 5, pp. 10–16, 2015.

[31] Y. Elkhatib, B. Porter, H. B. Ribeiro, M. F. Zhani, J. Qadir, and E. Rivière, "On using micro-clouds to deliver the fog," *IEEE Internet Computing*, vol. 21, no. 2, pp. 8–15, 2017.

[32] I. Farris, L. Militano, M. Nitti, L. Atzori, and A. Iera, "Federated edge-assisted mobile clouds for service provisioning in heterogeneous iot environments," in *Internet of Things (WF-IoT), 2015 IEEE 2nd World Forum on*. IEEE, 2015, pp. 591–596.

[33] A. R. Elias, N. Golubovic, C. Krintz, and R. Wolski, "Where's the bear?-automating wildlife image processing using iot and edge cloud systems," in *Internet-of-Things Design and Implementation (IoTDI), 2017 IEEE/ACM Second International Conference on*. IEEE, 2017, pp. 247–258.

[34] T. Taleb, A. Ksentini, and R. Jantti, "Anything as a Service for 5G mobile systems," *IEEE Network*, vol. PP, no. 99, pp. 12–19, Nov 2016.

[35] W. Felter, A. Ferreira, R. Rajamony, and J. Rubio, "An updated performance comparison of virtual machines and linux containers," *IBM Research Report*, July 2014.

[36] ETSI, "Network Functions Virtualisation - An Introduction, Benefits, Enablers, Challenges, Call for Action," Tech. Rep., Oct. 2012.

[37] ONF, "Software-defined networking: The new norm for networks," *ONF White Paper*, 2012.

[38] K. Kumar, J. Liu, Y.-H. Lu, and B. Bhargava, "A survey of computation offloading for mobile systems," *Mobile Networks and Applications*, vol. 18, no. 1, pp. 129–140, 2013.

[39] Y. Lin, T. Kämäräinen, M. Di Francesco, and A. Ylä-Jääski, "Performance evaluation of remote display access for mobile cloud computing," *Computer Communications*, vol. 72, pp. 17–25, December 2015.

[40] B. Kantarci and H. T. Mouftah, "Trustworthy sensing for public safety in cloud-centric internet of things," *IEEE Internet of Things Journal*, vol. 1, no. 4, pp. 360–368, 2014.

[41] S. R. Ellis, K. Mania, B. D. Adelstein, and M. I. Hill, "Generalizeability of latency detection in a variety of virtual environments," in *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, vol. 48, no. 23. SAGE Publications, 2004, pp. 2632–2636.

[42] R. T. Azuma, "A survey of augmented reality," *Presence: Teleoperators and virtual environments*, vol. 6, no. 4, pp. 355–385, 1997.

[43] M. Jarschel, D. Schlosser, S. Scheuring, and T. Hoßfeld, "An evaluation of QoE in cloud gaming based on subjective tests," in *Innovative Mobile and Internet Services in Ubiquitous Computing (IMIS), 2011 Fifth International Conference on*. IEEE, 2011, pp. 330–335.

[44] T. Kämäräinen, M. Siekkinen, Y. Xiao, and A. Ylä-Jääski, "Towards pervasive and mobile gaming with distributed cloud infrastructure," in *2014 13th Annual Workshop on Network and Systems Support for Games*. IEEE, 2014, pp. 1–6.

[45] D. Fitzgerald, N. Trakarnratanakul, L. Dunne, B. Smyth, and B. Caulfield, "Development and user evaluation of a virtual rehabilitation system for wobble board balance training," in *Engineering in Medicine and Biology Society, 2008. EMBS 2008. 30th Annual International Conference of the IEEE*. IEEE, 2008, pp. 4194–4198.

[46] C.-Y. Huang, K.-T. Chen, D.-Y. Chen, H.-J. Hsu, and C.-H. Hsu, "GamingAnywhere: The first open source cloud gaming system," in *Proc. 4th ACM Multimedia Systems Conf. (MMSys '13)*, Oslo, Norway, Feb.-March 2014.

[47] Y. Deng, Y. Li, R. Seet, X. Tang, and W. Cai, "The server allocation problem for session-based multiplayer cloud gaming," *IEEE Transactions on Multimedia*, 2017.

[48] W. Xia, P. Zhao, Y. Wen, and H. Xie, "A survey on data center networking (dcn): infrastructure and operations," *IEEE Communications Surveys & Tutorials*, vol. 19, no. 1, pp. 640–656, 2017.

[49] S. Dutta, T. Taleb, and A. Ksentini, "QoE-aware elasticity support in cloud-native 5G systems," in *IEEE Int'l. Conf. on Comm. (ICC 2016)*. IEEE, 2016, pp. 1–6.

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