IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received February 16, 2020, accepted March 3, 2020, date of publication March 9, 2020, date of current version March 18, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.2979385

Beyond First Impressions: Estimating Quality of Experience for Interactive Web Applications

HAMED Z. JAHROMI[1], DECLAN T. DELANEY[2], AND ANDREW HINES[1], (Senior Member, IEEE)

[1]School of Computer Science, University College Dublin, Dublin 4, Ireland
[2]School of Electrical and Electronic Engineering, University College Dublin, Dublin 4, Ireland

Corresponding author: Hamed Z. Jahromi (hamed.jahromi@ucdconnect.ie)

This work was supported in part by the Research Grant from the Science Foundation Ireland (SFI), and in part by the European Regional Development Fund under Grant 13/RC/2077 and Grant SFI/12/RC/2289_P2.

ABSTRACT The number of web applications for both personal and business use will continue to increase. The popularity of web applications has grown, increasing the need to estimate Quality of Experience for web applications (Web QoE). Web QoE helps providers to understand how their end-users perceive quality and point towards areas to improve. Waiting time has been proven to have a significant influence on user satisfaction. Most studies in the field of Web QoE have focused on modelling Web QoE for the user's first interaction with the application, e.g., the waiting time for the first page load to complete. This does not include a user's subsequent interactions with the application. Users keep interacting with the application beyond the first page load resulting in an experience that consists of a series of waiting times. In this study, we have chosen web maps as a use case to investigate how to measure waiting time for a user's interactions across a web browsing session, and to measure the correlation between waiting time and user-reported perceived quality. We provide a short survey of existing Web QoE estimation metrics and models. We then propose two new measures: interactive Load Time (iLT) and Total Completed interactive Load (TCiL) to establish the waiting time associated with a web application user's interactions. A subjective study confirms a logarithmic relationship for interactive web application sessions between iLT and perceived quality. We compare the correlation between QoE for iLT and the state of the art, non-interactive equivalent, Page Load Time (PLT)/Waiting Time. We demonstrate how the iLT/QoE fitting curve deviates from PLT/QoE. The number of clicks in completing tasks and TCiL are explored to explain the connections between user's interactions behaviour and the perceived quality.

INDEX TERMS Web QoE, interactive QoE, quality measurement, quality metrics, time metrics, waiting time, iLT, TCiL.

I. INTRODUCTION

Web-based application performance relies heavily on Quality of Service (QoS) metrics for optimised network parameters. These optimization techniques do not consider human factors and parameter adjustments may not result in perceptible quality improvements from a user's perspective [1]. Quality of Experience (QoE) considers user experience factors beyond the service [2], i.e., context, user, content and system factors (Figure 1). QoE provides further insights into the user's quality perception and their satisfaction [3], [4]. The relationship between QoE and QoS allows us to understand how to improve the end-to-end performance with respect to user satisfaction [5].

Web application quality can be measured at different levels: network, application, client and end-user (see Figure 2). Except for the end-user level, objective quality metrics (see Table 1) can be deployed to monitor QoS. Estimating end-user perceived quality requires subjective experiments which are expensive and time-consuming. In order to include the end user satisfaction in objective metrics, researchers have been working to develop mapping functions between QoS-based objective metrics and the subjective QoE experimental results [5].

These mapping functions can help identify the interdependence between factors such as network and

The associate editor coordinating the review of this manuscript and approving it for publication was Derek Abbott.

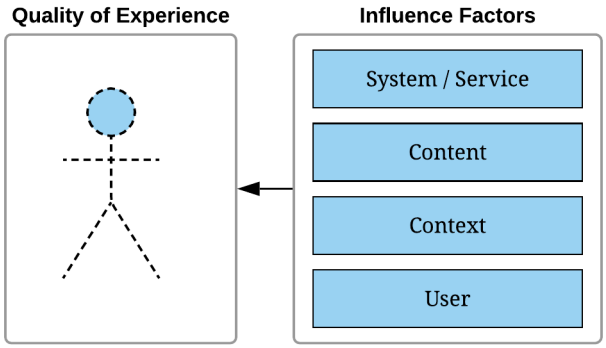


FIGURE 1. Illustration of quality of experience influence factors. Considering QoE of Web applications, System/Service factors include the network and computational components, Content refers to how the information is presented (i.e. media type, resolution, content length), Context covers the subject and the theme of the content (i.e. business, leisure, educational). Finally, The User covers the user's experience and expectations from the web application.

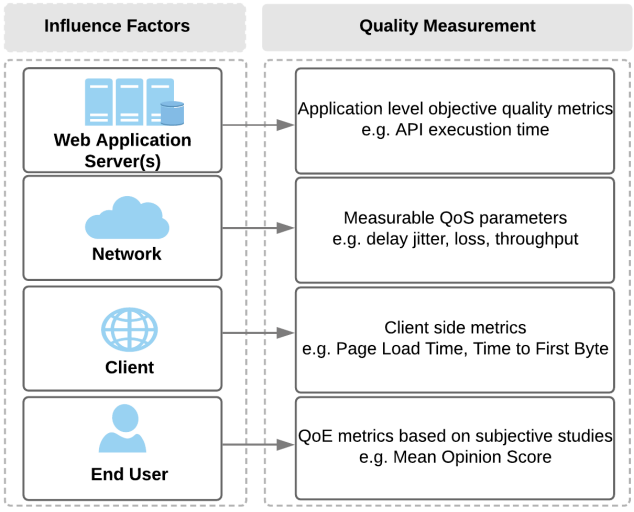


FIGURE 2. Illustration of quality measurement in the context of Web QoE. The mapping between quality metrics can help to understand the interdependence between factors.

application factors. For example, a mapping function between a network QoS metric and an application Key Performance Indicator (KPI) or application KPIs to experimental findings of subjective perceived quality [6], [7]. The mapping between metrics facilitates the development of objective metrics and models that explain the perceived quality (QoE) with a lower cost and overhead [5].

To date, an extensive amount of research has been carried out on the relationship between different influencing factors of QoE for Voice, Video and Multimedia applications [8], [9]. More recently, QoE has been applied to a broader range of applications including web-based applications (Web QoE) [4]. Web QoE refers to the quality of experience of web services that are based on the HTTP protocol and are accessed via a web browser [10]. Web shopping, downloading files or web mapping applications are familiar examples of such applications. Web applications follow a request-response paradigm in which the user makes a request, the server processes the request and issues a response. The network transports data between web server and user's browser. In contrast to audio- and video-based application QoE, where psycho-acoustic and psycho-visual are the most influential factors, the amount of time that the user waits to receive the information is a key factor impacting Web QoE [10], [11].

A common approach for Web QoE estimation is based on analysing the relationship between waiting time metrics and the end-user's perceived quality [4], [12]. Over time, web technologies have evolved to allow developers to create new generations of interactive and immersive web applications. For example, AJAX (Asynchronous JavaScript And XML) Push and HTML5 Websockets allow a web server to push data to a browser, without the browser explicitly requesting it. The user interacts with the web page and receives more content without making a specific request to retrieve a new page. The majority of the existing QoE estimation metrics, however, are defined based on the first page load and do not reflect the user's subsequent interactions across the browsing session (Figure 3). For clarity, the term "user interaction" is used in this paper to refer to a cycle, including a user's action and its response. An interaction is initiated when an action is made, and the user is waiting for a response. i.e. click on a linked object, submitting a search form, typing in a text-box with AJAX auto-complete feature.

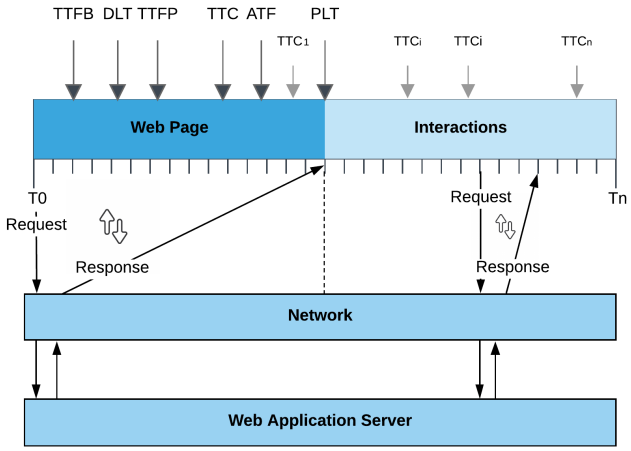


FIGURE 3. Illustration of quality measurement metrics. The majority of objective quality metrics are defined based on the first-page load. However, the web user's interactions continue to happen past the PLT.

In this paper, we investigate how to measure waiting time related to the web user's interactions beyond the first page load. This measurement needs to be meaningful to QoE and the perceived quality. We first provide a short survey and background information on QoE estimation metrics and models. We explore and categorise the existing objective time metrics. We continue by reviewing the Web QoE models and describe how the models use time metrics to estimate the perceived quality. Through a literature survey, we explain the evolution of web technologies and establish the need

for a time metric that covers users' interactions during a web browsing session. We then introduce *interactive Load Time* (iLT) and *Total Completed interactive Load* (TCiL) metrics to measure the waiting time associated with the user's interactions. We consider web mapping application as a use case and depict the test-bed and methodology used for the subjective study. The subjective study confirms the logarithmic relationship between iLT and the perceived quality. It follows by a correlation comparison between iLT and the non-interactive equivalent state of the art Page Load Time (PLT). We use TCiL to demonstrate that the fitting curve of iLT differs from PLT due to the continuous interactions of the users. By comparing TCiL and number of clicks, we show that TCiL provides more insights on the perceived quality in comparison to the number of clicks. TCiL can be utilised to understand the threshold of the tolerance of user's waiting time. Finally, we summarise and discuss the challenges involved in measuring iLT and TCiL, our thoughts and possible directions for future studies.

II. RELATED WORKS

Quality metrics and models are the key components of Web QoE analysis. Quality metrics are used to measure the efficiency and performance of Web applications. QoE models utilise objective quality metrics to estimate the perceived quality of the end-users [3], [13]. Perceived quality is widely measured in the domain of QoE using subjective user ratings where the subject rates quality on 5-point Absolute Category Rating (ACR) scale (1:Bad, 2:Poor, 3:Fair, 4:Good, 5:Excellent) [14]. Using the results from a group of test subjects, a Mean Opinion Score (MOS) is expressed as a single rational number, computed as the arithmetic mean of the subjective ratings.

Researchers utilise different variations of waiting time metrics for Web QoE analysis. In this section, we explore the time metrics available to measure waiting time for web applications. It is followed by a review of the existing Web QoE models that incorporate the time metrics.

A. WAITING TIME METRICS

Waiting time metrics in web applications are commonly used to estimate the user's satisfaction [15]. Web QoE researchers consider waiting time as an objective and measurable metric to build a mapping function between system/technical parameters and the subjective user's QoE. Web QoE is often simply estimated as a function of waiting time [13].

In Table 1 we have summarized the objective time metrics used in the Web QoE estimation models [16], [17]. Each metric measures the waiting time for a particular event occurring over the course of web application usage. The time metrics are divided into two distinct categories:

- **Time instant metrics:** Computed based on measuring the time instant of an event. For example, the time that a web page is completely loaded (PLT).
- **Time integral metrics:** Used to quantify how fast a web page is loaded by integrating all events of a given type tracked during the progress of a web page. For instance, the speed of loading a page starting from navigating to a URL until the time that browser has finished painting the visible part of the screen.

TABLE 1. The common time instant and integral metrics measured in seconds. The objective metrics are used to build a model for the evaluation of the perceived quality of web applications.

Time Instant Metrics	
Metric	**Description**
Time to First Byte (TTFB)	Time at which the first byte is received
Time to DOM Load (TDOM)	Time at which Document Object Model (DOM) is Loaded
Time to First Paint (TTFP)	Time at which the first pixel is painted on a screen
Time to First Contentful Paint (TTFCP)	Time at which the the browser first paint any object from the DOM, including any text, images, non-white canvas onto the page
Time to First Meaningful Paint (TFMP)	Time at which the significant portion of above-the-fold layout change has happened, and web tests have loaded
Time To Click (TTC)	Time at which the user performs the first click on an object
First Input Delay (FID)	Time at which the page has displayed the meaningful content but not yet interactive
Time To Interactive (TTI)	Time at which the page becomes interactive. "Interactive" is the point where the page has displayed the meaningful content, and the page responds to user interactions within 50 milliseconds
Above The Fold (ATF)	Time at which the Above-The-Fold content is rendered
Page Load Time (PLT)	Time from the start of navigation until the beginning of the window load event
Time Integral Metrics	
Metric	**Description**
SpeedIndex (SI)	Integral of complementary visual progress by calculating the mean pixel histogram difference of the current Web page at time t and a state of the page. SI can consider different time instant metrics as the state of the page. i.e. TTI, ATF and PLT
Perceptual SpeedIndex (PSI)	Utilises the Structured SIMilairy Index (SSIM) and calculates the integral of complementary visual progress
ByteIndex (BI)	Integral of complementary byte-level completion
ObjectIndex (OI)	Integral of complementary object-level completion

1) TIME INSTANT METRICS

Time instant metrics are simple to measure. They are used to measure the waiting time by tracking the amount of time the users wait until a particular event occurs. For instance, when a user first navigates to a web site, a TCP connection to a web server has to be made. The connection introduces a delay in data transport. As shown in Table 1, Time to First Byte (TTFB) is used to measure the time at which the client receives the first byte from the server. The browser finishes constructing the Document Object Model (DOM) at Time to DOM Load. The browser then starts rendering the web page at the Time to First Paint. At the Time of

First Contentful Paint, the first object has completed loading and rendering the information to the browser screen. It is followed by the Time to First Meaningful Paint which measures the time in which the browser has meaningfully painted the visible part of the browser on user's screen (current viewport). The browser completes painting the current viewport at the Above-The-Fold (ATF) time. Finally, it finishes loading all the visual and non-visual objects at the time of PLT.

Time to Click (TTC) is defined to measure the first interaction of the user. The event usually occurs between Time to First Meaningful Paint and PLT [13]. However, TTC may fail in measuring the responsiveness of a web page. For example, If the browser's main thread is still busy loading scripts, the page will no respond to the user's clicks. Time to Interactive (TTI) [18] tries to cover TTC's shortcomings by measuring how long it takes for a web page to become interactive. In [18] the term 'interactive' is characterised based on the following criteria:

- The page has already displayed meaningful content on the user's screen.
- All the applicable web elements are responsive. i.e. if there is a clickable object, all the event handlers associated with the object is loaded, and the user can click on the object.
- The page responds to the user interactions in 50 ms. i.e. if the user clicks on a button, the user can see the button is pressed in 50 ms.

TTI is a useful metric that shows the user experience from the interactivity point of view. For example, a web page might appear to be fully loaded, but the user is still unable to click on any object. Similarly, First Input Delay (FID) is a metric that measures the first impressions of a web page from an interactivity perspective [19]. While TTI measures how long it takes to become interactive, FID measures the delay that users experience when a meaningful content is displayed, but the web page is no yet interactive [18], [19]. However, these metrics are only effective in a page-by-page navigation model. In Section IV, we propose two new metrics that expands the quantification of user's interactivity beyond the first page load by separating the time measurement based on the user's actions.

Despite the prevalence of time instant metrics for the evaluation of the perceived quality, such metrics have proven shortcomings [13]. The most important limitation with time instant metrics is that the measured user experience depends only on the wait associated with retrieving and displaying the web page. However, during this process other events occur that a single time instant metric does not capture. In [13], [20], the authors demonstrated that estimating QoE using the same time instant metric for two different web applications can yield differing results. For instance, PLT is a useful metric to estimate QoE of a web page with contents limited to the current screen viewport (i.e. no scrolling required to see the full page). However, for a web page with extended content, ATF is more effective. To fill this gap and cover all the events in the web page waterfall, *Time Integral Metrics* have been proposed [21].

2) TIME INTEGRAL METRICS

In 2012, Google developed a time integral metric called Speed Index (SI) [21]. SI is a page load performance metric that represents how fast (in milliseconds) the visible parts of a web page are populated [20]. The lower the SI score is, the better the user's perception of performance. Time Integral Metrics use the following function to estimate the loading speed:

$$X = \int_0^{t_{\text{end}}} (1 - x(t))dt \tag{1}$$

where X is the value of the speed metric, t_{end} is the time the last event has happened, and $x(t) \in [0, 1]$ is the time evolution of the progress to reach t_{end}. For example, *ATF* can be defined as the t_{end} time while $x(t)$ is the completion ratio of the web page over time. The completion ratio of SI is calculated based on the Mean Pixel Histogram Difference (MPHD) between the current state of the web page at time t and the state of the page at the *ATF* time. The Perceptual Speed Index (PSI) uses Structured Similarity Index (SSIM) to compute the completion ratio of the web page [22].

Both SI and PSI use a series of snapshots (at a rate of 10 frames per second) from a web browsing session. The frames are analysed in the same sequence and time order to infer a visual completion fraction. The visual progress calculation is computationally expensive. Therefore, the researchers used the SI concept and proposed ByteIndex (BI) and ObjectIndex (OI) [20]. OI and BI use the browsers heuristics to estimate the loading speed. BI uses the ratio of byte completion as $x(t)$ starting from t_0 until t_{end} and OI uses the ratio of object completion as $x(t)$ starting from t_0 until t_{end}. BI and OI can consider different state of the page as t_{end} (i.e. ATF or PLT are commonly used as t_{end}).

If we look at the time integral metrics from an interactivity perspective of the first page load, the metrics can be bounded to TTI, TTC or FID and bring user's interactivity into account. However, the integral function of such metrics gives a lower weight to the delay occurred beyond ATF. Consequently, the measurement may not accurately quantify the speed of loading when a page is loaded but not yet responsive (the main thread of browser is still busy). Furthermore, The literature shows that the time integral metrics are difficult to measure and computationally expensive but outperform the predictive capability of time instant metrics for Web QoE estimation [13], [16].

B. WEB QoE MODELS

Much of the current literature studying Web QoE has focused attention on the development of PLT-based QoE models. In [23], the authors propose a generic QoE model where QoE and technical QoS metrics are correlated through an exponential relationship referred to as the IQX (exponential Interdependency of Quality of eXperience and QoS).

The authors evaluated the model for voice, video and web applications. Using waiting time as an objective metric leads to the following QoE IQX equation:

$$\mathrm{QoE}^{\mathrm{IQX}}(t) = \alpha e^{-\beta t} + \gamma. \tag{2}$$

where t is the waiting time measured by a time instant or time integral metric, α, β and γ are an empirically derived constants. The constants are tuned in accordance with the context (i.e Web, VOIP, Video). The authors illustrate that when the current level of QoE is high, a small variation in the QoS is perceptually noticeable. Thus, yielding to an exponential relationship between QoE and waiting time.

Egger *et al.* [24] developed a Web QoE estimation model based on the Weber-Fechner law, which is a human perception law drawing from the field of psychophysics [25]. In [24], the proposed Web QoE model is derived from a hypothesis called WQL. WQL assumes that the relationship between *W*aiting time and its *Q*oE evaluation on a linear ACR scale is *L*ogarithmic. The authors used the following fitting function and validated the WQL hypothesis:

$$\mathrm{QoE}^{\mathrm{WQL}} = a - b\ln(t) \tag{3}$$

where t refers to the waiting time measured by a time instant or time integral metric, a and b are derived by minimizing the least square errors between the fitting function and the MOS values. Egger *et al.* [24] state that while WQL is valid for simple waiting time transactions (e.g. PLT), it is not sufficiently sophisticated for use with interactive web applications. The authors explain that according to the Weber-Fechner Law, the perception of page load time deviates from the objectively measured page load time, i.e., the user may perceive a two seconds page load time as three seconds. Thus, the modelling of web browsing QoE have to be reviewed and redesigned.

In [24], [26], QoE is considered as a function of PLT. Both studies demonstrate that PLT has a significant effect on the user's overall Web QoE.

Gao *et al.* [22] explore different aspects of the web page loading process. The authors investigate the most influential factor in perceiving the loading speed. The authors studied the perceived loading performance of ATF for 500 websites. They ran a subjective study and presented two web-pages side-by-side using pairwise comparison to identify the faster page. They found that commonly used time instant metrics such as PLT and TTFB failed to accurately predict the end-user perception. Interestingly, their results show that the users can discriminate the speed of two web pages before the "VisualComplete" event (i.e. ATF). Gao *et al.* [22] conclude that in addition to PLT, the visual aspect of the web content also has a significant influence on the perceived loading speed.

In [13], the authors examine the relationship between Speed Index and Web QoE. They utilise an existing public dataset to establish the inter-dependency between SI and MOS values using the IQX and WQL models. The results show that the time integral metrics bounded to ATF are more effective than pure PLT as inputs used to estimate Web QoE [13].

C. MEMORY EFFECTS AND WEB QoE

Memory and recency effects are psychological phenomena which are important to Web QoE. Recency effects caused by short-term memory; it occurs when the latest perceived information is the most influential factor in human judgement [27].

In the field of Web QoE, researchers have been investigating the influence of memory and recency effect on the perceived quality. In [10], [28], the authors explored the influence of the users' psychological factors on Web QoE. They have conducted user studies to investigate how the users' previous experience, memory and recency effects have an impact on Web QoE. Their research shows the user memory influences the user perception of waiting times. In particular, the authors establish that in addition to the current level of QoS, quality of the last downloaded web page has a significant influence on the user's QoE (recency effect). Thus, the memory effect is a key influential factor impacting Web QoE.

In [10], the authors propose three different QoE models that include the implications of the memory effect and the time-dynamics of human perception into account. The authors have utilised Support Vector Machines (SVM), iterative exponential regressions, and two-dimensional Hidden Markov Models (HMM) to extended the basic QoE models and take memory effect into account. The authors also established that the memory effect steps down if the user experiences the same waiting time for several web pages in a row.

III. WEB EVOLUTION AND QoE CHALLENGES

Over the years, web applications have been transitioning from traditional models which use static pages to dynamic user interfaces with real-time and collaborative features (Figure 4). It is now common for web applications to link or embed information from other web applications such as

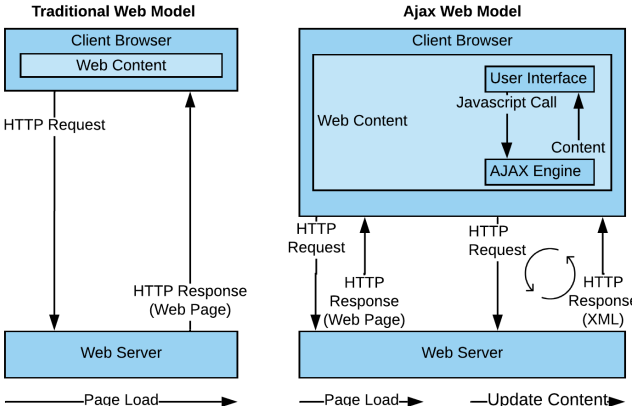


FIGURE 4. Illustration of traditional web model vs AJAX web model. In traditional web models, an HTML request results in a full page refresh. In an AJAX Web model, the user requests a new content using XHR request and the respective contents/objects will be retrieved and displayed dynamically (an in-place update).

social media or open data platforms [29]. This contrasts with the traditional web model that is based on a multi-page interface where every request results in a full page refresh.

AJAX and Comet programming are now commonly used in the development of interactive web applications [30]. These techniques eliminate the page-by-page navigation limitations of the traditional web model. A single-page Comet/AJAX application loads data and displays page components independently. A user action can result in a partial update of the web page while other page components remain visible. This is refered to as an in-place update. This influences the level of interactivity, responsiveness and user satisfaction [31]. The AJAX web model incorporates XML, JavaScript, HTTP and XHTML, combining together to facilitate asynchronous communication between client and web server.

Web mapping applications are an example of applications using AJAX, e.g. Google Maps has been integrated into geographic information system applications using AJAX [32]. The interactive nature of web mapping applications requires user requests to be quickly actioned in order to maintain relevance, flow experience and user attention [33]. Using AJAX allows web maps to load progressively, and each time a user re-centers the map (pans), some map sub-images (tiles) are kept for display while new tiles are fetched to update the view using XMLHttpRequest (XHR).

Although the evolution of web applications elevate the level of interactivity, responsiveness and user satisfaction, this has made web applications complex entities leading to more challenges for estimating the QoE [11]. The commonly used metric in QoE models (illustrated in Table 1) are based on the first-page load. However, in a single page interactive web application, the user keeps interacting with the applications and results in a series of XHR, Websocket or HTTP/2 Push transactions. There is a waiting time associated with these transactions. Therefore, it is challenging to represent the waiting time for the web QoE estimation. For example, in traditional web models there was a single PLT associated with the page load, but in AJAX web models objects are often dynamically generated/displayed during and after the PLT.

Web applications have a flow experience that spans the first load and subsequent interactions. Current time metrics do not adequately capture the impact of user interactions on perceived quality.

IV. MOTIVATION

A growing body of literature has investigated modelling Web QoE using time instant and integral metrics [4]. The models generally utilise the first page load metrics. However, there has been little quantitative analysis on measuring waiting time caused by the user's subsequent interactions and its impact on the user's QoE. As we explained in the previous sections, the critical aspects of our motivations can be summarised as follows:

- When a user navigates a web application, the initial page load occurs, then the user starts interacting with the page elements. The duration of interactions is proven to be a significant part of the total web browsing session time [34]. Thus, the user's interactions needs to be considered in the perceived quality estimation.
- Interactive web application use technologies such as AJAX, Web Sockets and HTTP/2. The user's interaction causes a waiting time that does not reload the entire page and is proportional to the main PLT or ATF. The interactive waiting time can be caused due to numerous reasons: (1) an XHR transaction in AJAX web applications, (2) HTTP/2 push messages or (3) a client-side navigation process (i.e., visualising an SVG image, loading a cached content or a computational process using client-side scripts). The current waiting time metrics are not covering interactive waiting time.
- SI and PSI metrics do not include browsing events. Instead, they are looking at how fast the ATF of the page gets painted to the browser. In an interactive web application, the user's interaction does not reload the entire page. Therefore, using SI to measure the loading speed, results in an inaccurate calculation of the visual progress. Furthermore, due to the partial update of the content, it is challenging to define when the ATF is completed.

These reasons motivated us to explore a set of non-computationally expensive metrics that can be used to predict Web QoE for a user's interaction. We introduce two new metrics associated with the user's interactions:

1) interactive Load Time (iLT).
2) Total Completed interactive Load (TCiL).

A. INTERACTIVE LOAD TIME (iLT)

Interactive Load Time (iLT) is the time taken to complete an interactive load starting from a user interaction (e.g. a mouse click) to the completed update of the web application display. It is a client-side metric that measures the waiting time caused by a single interaction of the user beyond the ATF. Each interaction of the web user may initiate n number of overlapping events (E). For example, when a web mapping application user instigates a single zoom action, multiple XHR requests will be sent to the webserver. The webserver then transfers the required new tiles and information. Importantly, some information may be loaded from the local cache. As shown in Figure 5, iLT covers all the events for an interaction. We compute iLT as

$$\text{iLT} = \max\{E_{n_{\text{et}}}\} - \min\{E_{n_{\text{st}}}\} \tag{4}$$

where $E_{n_{\text{et}}}$ is the time that data processing and visualisation for the event E_n of a particular interaction is completed (i.e. all the non-visual and visual elements of the interaction are loaded and rendered). $E_{n_{\text{st}}}$ is the time that the first event of the user's interaction is initiated (i.e. when the user clicks on an element). The *min* operator finds the first event based on the time that the events are started (st). The *max* operator looks at the events ending times (et) and finds the last event.

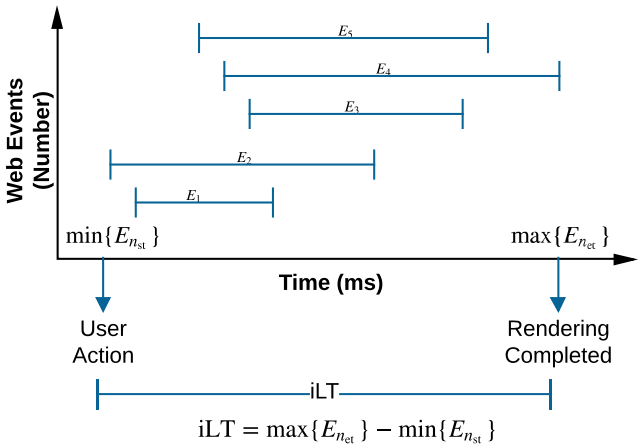


FIGURE 5. Illustration of iLT measurement based on the user's interaction and the web events. iLT is measured by computing the duration between starting time of the first event $E_{n_{st}}$ until the ending time of the last event $E_{n_{et}}$.

B. TOTAL COMPLETED INTERACTIVE LOAD (TCiL)

Total Completed interactive Load (TCiL) represents the number of times that the iLT is successfully measured for the entire session time. TCiL is computed based on Equation 5:

$$\text{TCiL} = \sum_{j=1}^{n} n_j \text{ where } n_j = \begin{cases} 1 & \text{if iLT}_{n_j} \in \mathbb{R} \\ 0 & \text{otherwise} \end{cases} \quad (5)$$

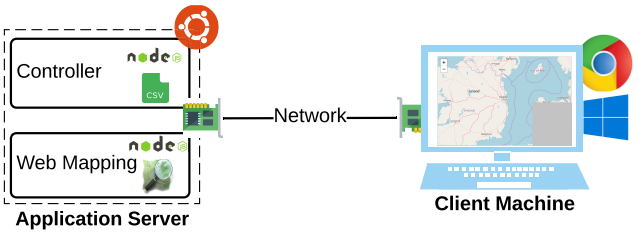
where n represents the total number of user interactions, j is a counter and n_j refers to an interaction number. The value of n_j will be one, if iLT has a measured value.

In a browsing session, the user will have multiple interactions, and an iLT will be measured for each interaction. However, the user may not wait for the responses to get completed before initiating the next interaction. This interrupts the response and prevents us from determining $E_{n_{et}}$ and a complete iLT measurement.

To investigate the effectiveness of the above metrics in explaining the relationship between interactive waiting time and the perceived quality, we have chosen to use a web mapping application as a representative use case. Many popular web mapping applications use raster image tiles to present a map view that can be zoomed or panned. When a user takes an action (i.e., zoom in, zoom out, pan or searching), it results in several XHR transactions to load the tiles. Some map tiles will be fetched from the server, and some tiles will be shown from the browser's cached objects. In this case, iLT refers to the amount of time that it takes for the map to load the map tiles in the client's browser.

In this research, by running a subjective experiment, we aim to establish a quantitative relationship between iLT and the user's QoE. The result of our study demonstrates how the relationship between iLT and QoE deviates from the state-of-the-art PLT. We also answer whether the iLT can accurately measure the waiting time caused by the user's interactions or not. Additionally, we compare TCiL and number of clicks and see which metric can better explain user's behaviour and experience.

Ultimately, we validate the generalisability of iLT and TCiL by looking at two different web mapping contents.



	Operating System
Server	Ubuntu 18.1
Client	Microsoft Windows 10 64 bit Home with Google Chrome 72.0.3626.109

FIGURE 6. The experimental platform developed for the user study. The platform simulated a real network environment and consisted of 2 node servers and a client machine. The network is dynamically manipulated to increase the iLT.

V. EXPERIMENTAL DESIGN AND PROCESS

To support further studies, the experimental platform and the data from this study have been shared.[1]

This subjective study has been designed in accordance with the general perceived performance estimation process defined in ITU-T Recommendation G.1030 [35] and ETSI (2010) Human Factors (HF); Quality of experience (QoE) requirements for real-time communication services [36].

A. STUDY PROCESS

Twenty eight individuals participated in this experiment. A pre-test questionnaire (Table 3) and 18 cases (9 using map tiles and 9 using satellite imagery tiles) were completed by each participant. Written instructions were provided for the predefined tasks (Section V-F). For each task, an expected iLT value is randomly selected and set from a group of nine waiting time values, by manipulating the network delay (Section V-B). The user experiences a series of iLTs for each task and, rates the perceived Web QoE for the task on 5-point ACR scale. Throughout the experiment, the main structure of the home page stays static, and only the map tile objects loaded. The participants were all members of the same educational institution and had similar prior experience using web mapping applications. This was considered an advantage in terms of the cohort homogeneity but it is acknowledged that it may also introduce other biases.

In the following subsections, we describe details of the experimental design parameters.

B. METRICS AND MEASUREMENTS

Table 2 presents the metrics captured in the user experiment. The web mapping application is instrumented to capture iLT,

[1] https://github.com/hzjahromi/iweb

TABLE 2. Objective and subjective metrics collected in the experiment. The metrics are collected at client side by instrumenting the web mapping application using JavaScript.

Objective		
Metric	**Unit**	**Description**
iLT	Seconds	iLT represents the amount of time taken for a tiled map to load all the required tiles fulfilling a single user action. For example, a mouse click to zoom in is an action that results in loading a set of map tiles.
TCiL	Count	Represents the number of times that all the map tiles are loaded before the user takes the next action. i.e. user may take the next action while the tiles are still blurry.
Number of Clicks	Count	Number of clicks made to perform a task
Subjective		
ACR	Category	The ACR scale is used to evaluate the perceived quality based on numbers that are assigned to the individual items, where Excellent equals to 5 and Bad equals to 1.

TABLE 3. Users characteristics questionnaire. The questionnaire is prepared using HTML form and filled by the participants prior to the subjective experiment.

User's Skills and Experience	
Questions	**Options**
Utilization of PC	Basic, Middle, Very good
Frequency of Using Web Maps	Daily, Weekly, Less than Monthly
Main Purpose of Using Web Maps	Commonly Used Functions (Navigation, Finding Locations, Tourism), Professional Functions (i.e Geospatial data, Business Specific)
User's Demography	
Which category includes your age	18-20, 21-29, 30-39, 40-49, 50-59, 60 or older
Gender	Male, Female, Other
Education	High school degree or equivalent, Bachelor degree, Graduate degree, Other
Eye Vision	Normal or Corrected Vision, other

TCiL and Number of Clicks. Upon the completion of each iteration, a page is presented with a 5-point ACR scale to the user to record the perceived quality of experience rating.

For each task the network delay was set to target iLT values close to those illustrated in Table 4, referred to as test cases. The choice of iLT values for the test cases are based on findings from previous web mapping and usability engineering studies [37], [38]. Based on the ITU-T recommendations, the sequence of test case evaluation is randomised for each user to minimise the interference between subsequent test cases and balance out the bias of memory effect [27], [39]. We have also considered stimulus spacing and frequency biases while selecting these values [40]. We expect that the selected values would result in an approximately normal distribution of judgments. The iLTs are achieved by imposing different network-level delays in the transport path,

TABLE 4. Targeted Values for different test conditions, we set network delay for each test case, executed the subjective task and collected the metric values. In this case the value of TCiL is fixed at ten in order to accurately measure each iLT. These values help us to compare the user's interactions/behaviours with these base values.

Map Content			
Test Case	**Mean iLT (s)**	**Sum iLT (s)**	**TCiL**
C1	0.10	0.90	10
C2	0.70	5.50	10
C3	2.80	22.00	10
C4	4.01	33.00	10
C5	6.50	50.50	10
C6	7.60	60.50	10
C7	10.05	87.50	10
C8	13.01	104.80	10
C9	14.13	113.05	10
Satellite Content			
C1	0.50	4.60	10
C2	1.30	10.00	10
C3	2.25	17.90	10
C4	4.60	36.70	10
C5	6.55	52.45	10
C6	10.00	79.35	10
C7	13.09	104.75	10
C8	17.15	137.25	10
C9	22.00	183.50	10

as described in [41]. The iLTs can approximately vary by 5% from the values shown in Table 4.

C. PLATFORM

We utilized a platform previously presented in [41] and added a feature to collect user quality ratings via a 5-point ACR scale. The platform has two components: a client machine and an application server. The application server hosts two services: a Web Mapping Application and a Controller. The controller manipulates the round trip network delay, captures the instrumented metrics and user ratings and stores them in a CSV formatted log file. The web mapping application provides a map of the world that contains a tiled map and collects the subjective quality rating. The main HTML page of the application is instrumented using a combined AJAX and JavaScript function to measure iLT, TCiL and Number of Clicks at the client side and passes the information to the controller. The controller stores the data and sets a new network delay for the next iteration.

We can adjust the iLT by instrumenting application or by changing network conditions such as increasing the network delay. Instrumentation of the application omits the common loading artefacts caused by the communication factors. Therefore, we used network delay to increase the iLT and keep the realistic map loading experience through progressive loading of tile images. Using network delay to vary the iLT allows us to investigate the realistic experience of the end user and be able to objectively reproduce a similar experience.

D. USERS AND CONTEXT

Participants were randomly recruited across different schools within a university context. We chose a scenario where end-users of the web mapping application explore a standard

mapping features in a non-emergency situation task that requires the user to find a location by searching, zooming and panning [42].

The user's skill-set (e.g. familiarity with the UI and functions) and user's objectives (e.g. pinpointing a place or navigating from one location to another) are understood to impact the user's experience [43]. Table 3 summaries the pre-test questionnaire undertaken to capture the cohort's previous experience with web applications as well as demographic information. Time perception between humans is more universal than familiarity with web mapping tools. Having a homogeneous cohort reduced the variation due to user experience.

We defined a single task with the same objective for all participants. Participants completed a paper-based pre-test training of the task steps to familiarise themselves with the procedures and the graphical user interface. This paper-based version with screenshots of the task was used in training in order to avoid biasing waiting time expectations.

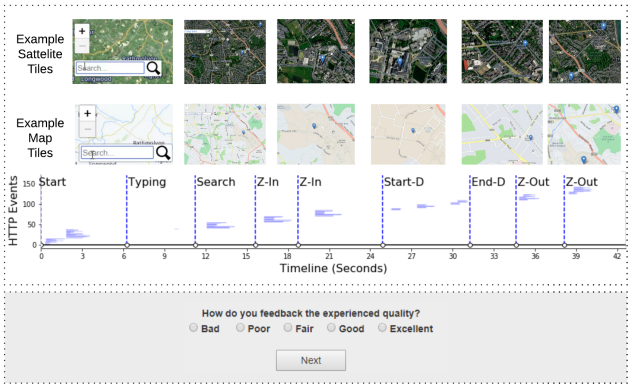


FIGURE 7. This graph illustrates an example of the experimental flow and the user's actions. The x-axis shows the task timeline and the y-axis represents the number of XHR events caused by the user's interactions. The dotted blue line shows the time that an action is taken and the horizontal blue bars represent the network traffic as a result of user's interactions. At the end of iteration, the user feedback the perceived quality.

E. WEB MAPPING CONTENT

To understand the effect of content on the perceived quality, we used two web mapping content types for testing: Map and Satellite Imagery. Satellite imagery map tiles have higher complexity and image density resulting in larger file sizes in comparison to map tile images.

F. TASK

A single task was defined to ensure all participants experience a similar level of interactivity. The goal is to search for a given location in a university campus and navigate the path to the nearby transport hub using common features of the web mapping services [42], [44]. To accomplish this, we designed a task with the following steps:

1) **Search:** Utilize the search box and look-up for university campus.
2) **Zoom In:** Once the university landmark is visible, zoom in three times to see the buildings and the street names.
3) **Navigate:** Pan to the left and find the main Road. Follow the main road to the north east until the transport hub is shown.
4) **Zoom Out:** Zoom out for three times to see both university and the transport hub.

G. USER SATISFACTION

For each iteration, the user rates the perceived quality upon the completion of the task using the 5-point ACR scale (as shown in Figure 7). For instance, for a test case, if performance quality from the web mapping application was as expected, it would be rated as excellent.

H. PHYSICAL ENVIRONMENT

To minimise the environmental error and increase the attention of the participants during the test, the participants completed the test in a controlled laboratory environment. The test was carried in a quiet lab room within a normal office setup where participants were asked to sit facing a 23'' PC monitor (Dell 2313H, 1920 $\times$ 1080 px) with a viewing distance of approximately 60 cm.

I. THE TEST USERS DEMOGRAPHICS AND EXPERIENCE

The information fields collected in the pre-test questionnaire completed by participants is summarised in Table 3. We validated our assumptions with respect to the homogeneity of the test cohort's technical experience and expertise with web mapping applications. The experiment was completed by 28 participants, 78% were male, and 22% were female, the age of participants ranges from 21 to 50 years old, with an average age of 33 years old. 18% and 72% of participants had Middle and Very Good level of skills with the utilization of computers, respectively. Regarding the familiarity with web maps, 89% of participants use web maps on weekly or daily basis, and 11% use web maps occasionally. The participants engaged in this study had graduate degrees. All participants had a normal or corrected vision.

J. COLLECTED DATA

We collected 479 data records, each containing the objective and subjective measures (Table 1). We excluded incomplete records that did not have a corresponding subjective rating and records where the iLT was not measured at least once for that particular iteration. This can occur for a variety of reasons as the iLT may not be measured if the participant does not wait for the content to be fully loaded while following the steps. For example, initially, the user zooms in but then starts panning before all the tiles have loaded. This prevents us from measuring the iLT for the zoom in action and if it occurs for all actions, the mean iLT value of the iteration will be zero. If we include a record with mean $iLT = 0\ ms$ in the data-set, it will reduce the accuracy and validity of our correlation analysis between iLT and the perceived quality.

K. INTRA-RATER AND INTER-RATER RELIABILITY

To quantify the reliability of our subjects (raters), we have considered two different types of reliability of the user study: intra-rater and inter-rater reliability. Intra-rater reliability shows to what extent the ratings of an individual user are consistent. inter-rater reliability measures to what extent the participants agree when rating the same set of test cases. We have computed intra-rater and inter-rater reliability for map and satellite separately.

We have used the methodology explained in [45], and utilised Spearman's Rank Correlation Coefficient to quantify intra-rater reliability. The correlation coefficient shows whether the relationship between two variables (iLT and ACR) can be explained with a monotone function. The ranking is expected to change proportionally with the value of the iLT, resulting in a different score for each test-case. If the ranking has no repetition, the result of the correlation will be one. The result of intra-rater reliability shows 0.68 and 0.78 for the map and satellite contents, respectively.

We have used the average measure Intra-class Correlation Coefficient (ICC) to measure inter-rater reliability. ICC shows the reliability of multiple raters averaged together. A higher ICC value indicates better inter-rater reliability. i.e. ICC of 1 illustrates a perfect agreement among the raters and 0 means a random agreement. The result of ICC for the map and satellite cases are 0.92 and 0.96, which falls under the acceptable inter-rater reliability threshold [46].

VI. ESTABLISHING THE RELATIONSHIP BETWEEN iLT AND PERCEIVED QUALITY

Each collected data record contains a series of iLTs and a quality rating on a 5-point ACR scale. From the data we can compute a Mean Opinion Score (MOS) and Mean iLT. MOS is computed as:

$$\mathrm{MOS}_j = \frac{\sum_{i=1}^{N} m_{ij}}{N} \tag{6}$$

where m_{ij} refers to the score by subject i for test case j and N is the number of the participants. Mean iLT is computed using the arithmetic mean over all individual Mean iLT of each test case (mean of means) based on:

$$\mathrm{iLT}_{pq} = \frac{\sum_{n=1}^{k} \mathrm{iLT}}{k} \tag{7}$$

and

$$\text{Mean iLT} = \frac{\sum_{p=1}^{N} \mathrm{iLT}_{pq}}{N}, \tag{8}$$

where iLT_{pq} refers to the mean of k number of iLTs experienced by subject p for the test case q. Mean iLT is the mean of iLT_{pq} based on the number of participants, N, for a test case q.

Figure 8 shows the relationship between mean iLT and the perceived quality using a MOS scale. The figure plots two sets of results for different content types: one for satellite imagery and one for simple maps. There are nine data points per experiment with 95% Confidence Interval (CI) error bars.

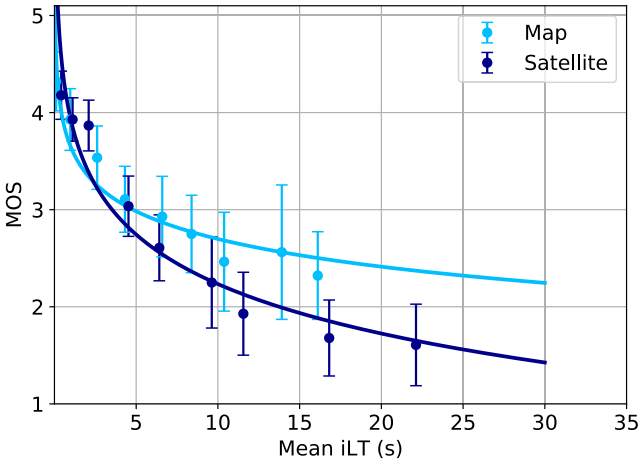


FIGURE 8. MOS Vs iLT with error bars indicating 95% CI. This figure represents the logarithmic relationship between MOS and iLT for both Map and Satellite imagery content. The *a* and *b* curve fitting parameters of Equation 9 are [−0.41,6.48] and [−0.73,9.00] for the map and satellite respectively. We have also computed the matching parameters *a* and *b* with more indicative bases than natural logarithm. The computed *a* and *b* parameters of Map content with base 2 and base 10 are [−0.28,6.48] and [−0.94,6.48] respectively. Similarly, For the satellite content, [−0.50,9.00] and [−1.69,9.00] are the computed *a* and *b* values with base 2 and base 10 respectively.

The fitting function, similar to WQL [24], is motivated by Weber-Fechner law stating that the perceived intensity is proportional to the logarithm of the stimulus [47].

As shown in Equation 9, we used a logarithmic fitting function and plotted as bold lines in Figure 8. However, as we did not ask the subjects to rate the QoE per action within the test sequences, the proposed model does not consider short term memory effects.

$$\mathrm{QoE}^{\mathrm{iLT}} = a \cdot \ln(t) + b \tag{9}$$

where a and b are coefficients derived using the following least square fitting functions:

$$a = \frac{\sum_{i=1}^{n} y_i - b \sum_{i=1}^{n} (\ln x_i)}{n} \tag{10}$$

$$b = \frac{n \sum_{i=1}^{n} (y_i \ln x_i) - \sum_{i=1}^{n} y_i \sum_{i=1}^{n} \ln x_i}{n \sum_{i=1}^{n} (\ln x_i)^2 - (\sum_{i=1}^{n} (\ln x_i))^2} \tag{11}$$

where t refers to the mean iLT.

From the results presented in Figure 8, we can see that a logarithmic fitting of $\mathrm{QoE}^{\mathrm{iLT}}$ can be used for both satellite imagery and simple map content. This logarithmic relationship corresponds with a well-known WQL hypothesis which shows iLT holds the expected relationship with QoE [24], [48].

We have used the coefficient of determination (known as "R-squared") to assess how good the fitting curve explains and predicts future outcomes. The computed R-squared for the map and satellite contents are 0.92 and 0.94, respectively. i.e. the R-squared of 0.92 for the map means that 92% of the dependent variable (MOS) is predicted by the independent variable (iLT).

The deviation in the fitting curve, however, shows that change in content type impacts the perception of waiting time and overall quality. The high-quality satellite imagery content impacts the users' quality perception for higher MOS levels (MOS $>$ 3). Additionally, we can see that while the iLT associated with both satellite and map content were similar, the participants' QoE ratings followed a higher average trend for the satellite over simple map content. This shows that the content impacts the user's expectations and quality perception [10], [11]. We see that when the map and satellite iLTs intersect at approximately three seconds, the map tiles yield better quality scores. This again highlights the importance of content as a Web QoE factor. In Section VII, we will further explore the impact of user interactions on the shape of the iLT curves.

VII. COMPARING iLT TO STATE OF THE ART NON-INTERACTIVE METRIC

In this section, we compare iLT, our proposed interactive Web QoE measure, with the non-interactive Web QoE metric, Page Load Time (PLT). This will allow us to compare how user interactions occurring after an initial interaction can impact Web QoE.

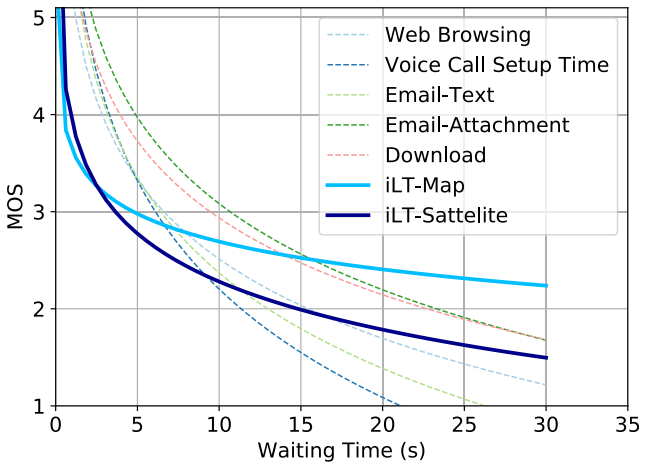


FIGURE 9. Comparison between iLT for web mapping application vs State-of-the-Art applications PLT/Waiting Time. The applications can be divided into two categories. The first category involves e-mail transmission with attached files and content downloads. The second category involves the time required to display the home page when accessing a web site (PLT), the time required to connect a voice call, and the time required to transmit a plain-text e-mail [48].

Figure 9 presents user waiting time (seconds) plotted against perceived quality (MOS) for a variety of web applications. This data was collected in several experiments and used to explore the relationship between users' QoE and waiting time for web page, voice, email and file-download web tasks. The dashed lines show results previously presented in [48]. The solid lines show the map and satellite mean iLT results from this work overlaid for comparison. The y-axis is the perceived quality on the MOS scale ranging from 1 to 5, and the x-axis represents the waiting time (Seconds).

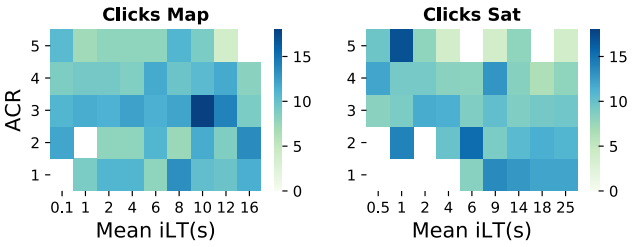


FIGURE 10. User behaviour: the relationship between interactive load time (mean iLT) and perceived quality (5 ACR levels) based on the number of clicks (the colour corresponds to the number of occurrences). Plotted by content: map tile (left) and satellite tiles (right).

Similar to the other web applications, we can see that the map and satellite mean iLT has a logarithmic relationship with the perceived quality. However, the different slopes observed for the iLT curves compared to non-interactive PLT curves point to the effect that subsequent user interactions have on the perceived quality. When a user interacts with a web application, it is expected that the interaction follows a "smooth flow" experience. The in-place (partial) update of an interactive web application can improve the "flow" experience. This could further explain the deviation in the fitting curves between Mean iLT and non-interactive application PLT.

VIII. USER BEHAVIOUR, WAITING TIME AND PERCEIVED QUALITY

How does user behaviour change based on waiting time and how is this related to perceived quality? We examine user behaviour using two metrics, first, the Number of Clicks that a user makes interacting with the web mapping application and second, the Total Completed interactive load (TCiL).

Figure 10 presents two heat maps that show the relationship between interactive load time (mean iLT) and perceived quality (5 ACR levels) from the experiments using map and satellite tiles respectively. The colour ranges from white to dark blue, indicating an increase in user activity, based on the number of clicks.

The top left quadrant is fast/high quality, and the bottom right is slow/low quality. The number of clicks varies across the range and we did not observe any trends in this data that provides insights into a relationship between waiting time, perceived quality and a change in the behaviour of the user in terms of their interaction as measured by number of clicks.

Similarly, Figure 11 presents two heat maps that show the relationship between interactive load time (mean iLT) and perceived quality (5 ACR levels) from the experiments using map and satellite tiles respectively. In Figure 11 the colour ranges from white to dark blue, indicate an increase in the number of completed interactions by the user, by counting the number of completed interactive load (TCiLs) recorded. If a user clicks to initiate a new request to the application before the page load has competed for the previous interaction a TCiL is not recorded.

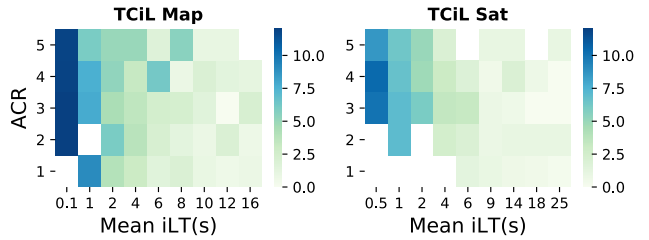


FIGURE 11. User behaviour: the relationship between interactive load time (mean iLT) and perceived quality (5 ACR levels) based on the number of completed interactions, TCiLs. The colour corresponds to the number of interactions completed. Plotted by content: map tile (left) and satellite tiles (right).

The top left quadrant is fast/high quality, and the bottom right is slow/low quality. Looking at the satellite results (right hand plot), the trend shows that as quality increase and waiting time decrease, the number of completed interactions increases.

The first column from the left of the TCiL for Sat subplot shows that, for iLT $<$ 0.5 seconds, the number of completed loads varies from eight to 14 times. For the iLTs between 0.5 and two seconds, the number of completed loads trends higher that for iLT $>$ 2 seconds. We postulate that users exhibit patience and are willing to wait for the result of each interaction for iLTs $<$ 2 seconds. This finding for the web interactions is consistent with the *two seconds* rule observed in [49] for the text entry and editing tasks on computer terminals.

Taken in the context of the QoE factors introduced in Figure 1, user experiences will have pre-established expectations regarding the time required to complete an interaction. If the iLT associated with an interaction faster than expected, the user will be satisfied with the quality. If the iLT is perceived as slower than expected, the QoE will be lower and they may start loosing attention and focus on the task as they become consciously aware of their dissatisfaction. According to [49], users are satisfied with response times of less than a second for most tasks, however, a two seconds response time is generally acceptable and does not significantly impact the user's attention [38], [49].

Although the TCiL provides insight into tolerance thresholds for waiting time, it is apparent from Figure 11 that TCiL and perceived quality are not highly correlated. If they were, we would expect to see a smooth colour transition from a dark to light. The variation in TCiL highlights the challenge of measuring iLTs when the iLT goes beyond the user's tolerance thresholds. If the user loses patience and issues a new request, the load is incomplete and a new interaction begins. As a result, objective measurement of iLT is challenging as the instrumented metrics do not capture the whole experience.

IX. iLT AND USER INTERACTIONS

In the previous sections, we explored how the service (network), human (action) and content (sat/map) factors impact QoE (Figure 1). In this experiment, by normalising the network delay to zero and looking at the change in iLT per action, we will observe the differences in iLT for different user's actions. We will also demonstrate that iLT varies for the different content type. Thus, this highlights that the network delay is not the only influencing factor, and depending on the type of interactions and the content, the QoE will vary.

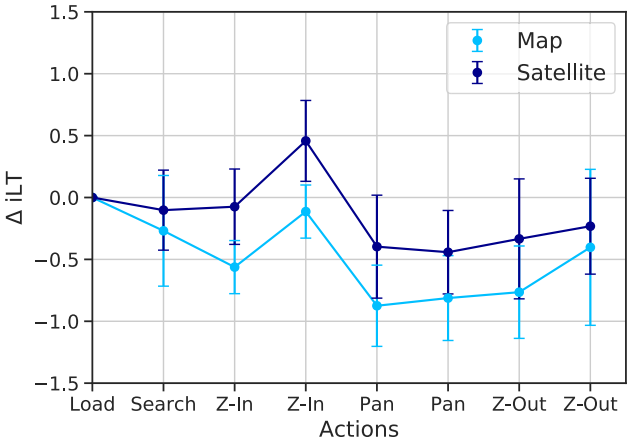


FIGURE 12. Illustration of difference in iLTs (ΔiLT) for different actions while excluding the effect of network delay. ΔiLT is normalised between −1 and 1 by considering relative to the first load time (iLT Load) as zero.

This experiment illustrates the value of capturing iLT per user's action rather than as a single session waiting time. This study is performed with seven participants (i.e. N = 7) invited from the same cohort (see Section V-I). For illustrative purposes, this cohort size was sufficient to illustrate the trends with statistical significance. The experiment focuses on differences in measured iLT and its relationship with the current action rather than perceived quality. We used the same methodology and process explained in Section V and instructed the users to:

- Execute the subjective task for 18 iterations (Table 4).
- Follow the task steps and wait for the result of each action before taking the next action This helps us to measure iLT accurately as it ensures that all interactions are completed.
- No quality rating is required at the end of each iteration.

We collected 1008 data records. Each record contains an iLT corresponding to a user's action. As explained in Section V, a mean iLT value is targeted for each test case. To better understand the difference in iLTs associated with actions, we have excluded the network effect for different test cases. The network effect was excluded by considering the first iLT (initial Load) as point zero. We then computed the difference between initial Load mean iLT and other actions mean iLT and normalised the values between −1 and 1. We will refer to this as ΔiLT. Figure 12 presents ΔiLT for sequential user actions. Results for both map and satellite content are plotted. The error bars are 95% CI of ΔiLT based on the user actions. If the ΔiLT of a user action is zero then the interaction load time corresponded with the initial load time. If the ΔiLT is negative, the content loaded faster than

TABLE 5. Mixed-model ANOVA results for fixed (F-Test) and random effects (Likelihood-Ratio Test). target variable: ΔiLT. Asterisks indicate levels of statistical significance. The *F* is the ratio of the mean-square value and χ^2 refers to the chi-square test.

Fixed	F	p	
Case	1.776	0.090	
Action	324.835	<0.001	***
Content	294.521	<0.001	***
Case:Action	5.719	<0.001	***
Case:Content	0.784	0.617	
Action:Content	50.853	<0.001	***
Random	χ^2	p	
Case:Userid	0.000	1.000	
Action:Userid	1.091	0.296	
Content:Userid	0.000	1.000	
Case:Action:Userid	0.000	1.000	
Case:Content:Userid	126.216	<0.001	***
Action:Content:Userid	3.321	0.068	
Userid	0.000	1.000	

*** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$

the initial Load while a positive ΔiLT means the content took longer to load than the initial load.

From our Mixed Model ANOVA analysis (Table 5) and Fig. 12 we can observe:

- If the network transportation delay was the only influencing factor, then the ΔiLT of all actions would be close to zero.
- The ΔiLT for the *Search* action is close to zero, i.e. it is similar to the *Load* action. The user has minimal interaction while taking the search action (just a click). For the Zoom and Pan actions, the user interaction with the application is high resulting in significantly variation in ΔiLT for these actions. Furthermore, the influence of users' actions on ΔiLT is statistically significant (Action, $p < 0.001$) which exerts that the iLT changes depending on the users' action.
- ΔiLT of map content differs from satellite content. This difference shows that the content is a factor influencing user interactivity (Action:Content, $p < 0.001$).
- Negative ΔiLT scores highlight that many actions competed faster than the initial load due to the fact that requests could rely on AJAX and caching to at least partially deal with the requests locally without needing a full data refresh from the server.

X. LIMITATIONS

In these experiments, users rated session QoE for a series of iLTs. The influence of memory effects (recency or primacy) on the perceived quality was not explored, i.e. are some iLT in the series more important to the rated QoE? The current study is a task-driven QoE study where the type of task and content could be salient factors influencing Web QoE. Our choice of a client-side metrics have been shown to be effective in measuring the user experienced waiting time for a session, but it is challenging to implement and measure them as a monitoring tool ''in the wild''. Lastly, we have used a web mapping application as a use case and representative of an interactive application that requires a smooth flow experience to develop the metrics and model. However, further evaluation of interactive web QoE is required to explore how the result and derived fitting curve generalise to a range of other web applications, contents and contexts.

XI. CONCLUSION

In this paper, we reviewed the existing models and metrics used for web QoE estimation, i.e. time instant and time integral metrics. Existing metrics are based on the first page load occurrence. We discussed why these current metrics are not sufficient to measure perceived quality for interactive web applications. We explained that the user keeps interacting with the application after the first page load, rendering PLT inadequate for interactive web applications. We introduced iLT and TCiL which capture waiting time associated with the user interaction. The iLT and TCiL metrics are computationally simple and both metrics can be continuously measured beyond PLT and ATF completion times. Using a subjective experimental study, we demonstrated that iLT has the same logarithmic relationship with user satisfaction as PLT. It also aligns with the well known WQL hypothesis. This confirmed that iLT is effective for Web QoE estimation. We showed that the slope of the fitting curve for mean iLTs/MOS deviates from the PLT/MOS and speculate that this is as a result of multiple factors including user interaction and the content type.

An investigation using total completed interactive load (TCiL) established that web users do not necessarily wait for a complete result of an interaction before taking the next action. This finding shows that measuring iLT can be complicated when a user interrupts a request as a result of the user's waiting time tolerance threshold. Further investigation of how to account for interrupted loads in low QoE scenarios could enhance the value of iLT as a metric. We believe that to have an effective Web QoE estimation metric for an interactive web application, the time integral metrics need to be re-designed and be able to capture the user's interactions beyond ATF and PLT. i.e. ByteIndex can be bounded to the time iLT start and endpoints.

REFERENCES

[1] B. Upadhyaya, Y. Zou, I. Keivanloo, and J. Ng, "Quality of experience: What end-users say about Web services?" in *Proc. IEEE Int. Conf. Web Services*, Jun. 2014, pp. 57–64.
[2] P. Le Callet, S. Möller, and A. Perkis, "Qualinet white paper on definitions of quality of experience," Eur. Netw. Qual. Exper. Multimedia Syst. Services, Lausanne, Switzerland, Tech. Rep. COST Action IC 1003, Mar. 2013.
[3] M. Alreshoodi and J. Woods, "Survey on QoE\QoS correlation models for multimedia services," *Int. J. Distrib. Parallel Syst.*, vol. 4, no. 3, p. 53, 2013.
[4] S. Baraković and L. Skorin-Kapov, "Survey of research on quality of experience modelling for Web browsing," *Qual. User Exper.*, vol. 2, no. 1, p. 6, Dec. 2017.
[5] D. Tsolkas, E. Liotou, N. Passas, and L. Merakos, "A survey on parametric QoE estimation for popular services," *J. Netw. Comput. Appl.*, vol. 77, pp. 1–17, Jan. 2017.

[6] H. Z. Jahromi, A. Hines, and D. T. Delanev, "Towards application-aware networking: ML-based end-to-end application KPI/QoE metrics characterization in SDN," in *Proc. 10th Int. Conf. Ubiquitous Future Netw. (ICUFN)*, Jul. 2018, pp. 126–131.

[7] H. Z. Jahromi and D. T. Delaney, "An application awareness framework based on SDN and machine learning: Defining the roadmap and challenges," in *Proc. 10th Int. Conf. Commun. Softw. Netw. (ICCSN)*, Jul. 2018, pp. 411–416.

[8] L. Skorin-Kapov, M. Varela, T. Hoßfeld, and K.-T. Chen, "A survey of emerging concepts and challenges for QoE management of multimedia services," *ACM Trans. Multimedia Comput., Commun., Appl.*, vol. 14, no. 2, pp. 1–29, May 2018.

[9] R. Huang, X. Wei, L. Zhou, C. Lv, H. Meng, and J. Jin, "A survey of data-driven approach on multimedia QoE evaluation," *Frontiers Comput. Sci.*, vol. 12, no. 6, pp. 1060–1075, Dec. 2018.

[10] T. Hoßfeld, S. Biedermann, R. Schatz, A. Platzer, S. Egger, and M. Fiedler, "The memory effect and its implications on Web QoE modeling," in *Proc. 23rd Int. Teletraffic Congr.*, Sep. 2011, pp. 103–110.

[11] S. Egger, T. Hossfeld, R. Schatz, and M. Fiedler, "Waiting times in quality of experience for Web based services," in *Proc. 4th Int. Workshop Qual. Multimedia Exper. (QoMEX)*, Jul. 2012, pp. 86–96.

[12] M. Lycett and O. Radwan, "Developing a quality of experience (QoE) model for Web applications," *Inf. Syst. J.*, vol. 29, no. 1, pp. 175–199, Jan. 2019.

[13] T. Hobfeld, F. Metzger, and D. Rossi, "Speed index: Relating the industrial standard for user perceived Web performance to Web QoE," in *Proc. 10th Int. Conf. Qual. Multimedia Exper. (QoMEX)*, May 2018, pp. 1–6.

[14] G. Albaum, "The Likert scale revisited," *Market Res. Soc. J.*, vol. 39, no. 2, pp. 1–21, Mar. 1997.

[15] M. Fiedler, P. Arlos, T. A. Gonsalves, A. Bhardwaj, and H. Nottehed, "Time is perception is money—Web response times in mobile networks with application to quality of experience," in *Performance Evaluation of Computer and Communication Systems. Milestones and Future Challenges* (Lecture Notes in Computer Science), vol. 6821, K. A. Hummel, H. Hlavacs, and W. Gansterer, Eds. 2010. Berlin, Germany: Springer, 2011.

[16] D. N. da Hora, A. S. Asrese, V. Christophides, R. Teixeira, and D. Rossi, "Narrowing the gap between QoS metrics and Web QoE using above-the-fold metrics," in *Passive and Active Measurement* (Lecture Notes in Computer Science), vol. 10771, R. Beverly, G. Smaragdakis, and A. Feldmann, Eds. Cham, Switzerland: Springer, 2018.

[17] A. Saverimoutou, B. Mathieu, and S. Vaton, "A 6-month analysis of factors impacting Web browsing quality for QoE prediction," *Comput. Netw.*, vol. 164, Dec. 2019, Art. no. 106905.

[18] *Time to Interactive*. Accessed: Nov. 10, 2019. [Online]. Available: https://developers.google.com/web/tools/lighthouse/audits/time-to-interactive

[19] *First Input Delay*. Accessed: Nov. 10, 2019. [Online]. Available: https://developers.google.com/web/updates/2018/05/first-input-delay

[20] E. Bocchi, L. De Cicco, and D. Rossi, "Measuring the quality of experience of Web users," *ACM SIGCOMM Comput. Commun. Rev.*, vol. 46, no. 4, pp. 8–13, Dec. 2016.

[21] *Speed Index (SI)—Webpagetest Documentation*. Accessed: Nov. 10, 2019. [Online]. Available: https://sites.google.com/a/webpagetest.org/docs/using-webpagetest/metrics/speed-index

[22] Q. Gao, P. Dey, and P. Ahammad, "Perceived performance of top retail Webpages in the wild: Insights from large-scale crowdsourcing of above-the-fold QoE," in *Proc. Workshop QoE-Based Anal. Manage. Data Commun. Netw. (Internet QoE)*, 2017, pp. 13–18.

[23] M. Fiedler, T. Hossfeld, and P. Tran-Gia, "A generic quantitative relationship between quality of experience and quality of service," *IEEE Netw.*, vol. 24, no. 2, pp. 36–41, Mar. 2010.

[24] S. Egger, P. Reichl, T. Hoßfeld, and R. Schatz, "'Time is bandwidth'? Narrowing the gap between subjective time perception and quality of experience," in *Proc. IEEE Int. Conf. Commun. (ICC)*, Jun. 2012, pp. 1325–1330.

[25] S. Dehaene, "The neural basis of the Weber–Fechner law: A logarithmic mental number line," *Trends Cognit. Sci.*, vol. 7, no. 4, pp. 145–147, Apr. 2003.

[26] M. Varela, L. Skorin-Kapov, T. Maki, and T. Hosfeld, "QoE in the Web: A dance of design and performance," in *Proc. 7th Int. Workshop Qual. Multimedia Exper. (QoMEX)*, May 2015, pp. 1–7.

[27] D. S. Hands and S. E. Avons, "Recency and duration neglect in subjective assessment of television picture quality," *Appl. Cognit. Psychol., Off. J. Soc. Appl. Res. Memory Cognition*, vol. 15, no. 6, pp. 639–657, 2001.

[28] J. Shaikh, M. Fiedler, P. Paul, S. Egger, and F. Guyard, "Back to normal? Impact of temporally increasing network disturbances on QoE," in *Proc. IEEE Globecom Workshops (GC Wkshps)*, Dec. 2013, pp. 1186–1191.

[29] A. Rio and F. B. E. Abreu, "Web systems quality evolution," in *Proc. 10th Int. Conf. Qual. Inf. Commun. Technol. (QUATIC)*, Sep. 2016, pp. 248–253.

[30] J. J. Garrett, "Ajax: A new approach to Web applications," *Adapt. Path*, pp. 1–3, Feb. 2005.

[31] A. Mesbah and A. Deursen, "An architectural style for Ajax," in *Proc. Work. IEEE/IFIP Conf. Softw. Archit. (WICSA)*, Jan. 2007, p. 9.

[32] A. Sayar, M. Pierce, and G. Fox, "Integrating AJAX approach into GIS visualization Web services," in *Proc. Adv. Int. Conf. Telecommun. Int. Conf. Internet Web Appl. Services (AICT-ICIW)*, 2006, p. 169.

[33] B. Veenendaal, M. A. Brovelli, and S. Li, "Review of Web mapping: Eras, trends and directions," *ISPRS Int. J. Geo-Inf.*, vol. 6, no. 10, p. 317, 2017.

[34] A. Edmonds, R. W. White, D. Morris, and S. M. Drucker, "Instrumenting the dynamic Web," *J. Web Eng.*, vol. 6, no. 3, pp. 244–260, 2007.

[35] *Estimating End-to-End Performance in IP Networks for Data Applications*, document G.1030, ITUT Recommendation, Geneva, Switzerland, 2005. [Online]. Available: https://www.itu.int/rec/T-REC-G.1030/en

[36] *Human Factors (HF); Quality of Experience (QoE) Requirements for Real-Time Communication Services*, document ETSI TR 102 643, 2010. [Online]. Available: https://www.etsi.org/deliver/etsi_tr/102600_102699/102643/01.00.02_60/tr_102643v010002p.pdf

[37] M. Fiedler, C. Eliasson, P. Arlos, S. Eriksén, and A. Ekelin, "Quality of experience and quality of service in the context of an Internet-based map service," Blekinge Inst. Technol., Karlskrona, Sweden, Tech. Rep. 1, 2008. [Online]. Available: https://www.iis.se/docs/74._Slutrapport_Report_Map_Service_Project_V1.1.doc

[38] J. Nielsen, *Usability Engineering*. Amsterdam, The Netherlands: Elsevier, 1994.

[39] *Methodology for the Subjective Assessment of the Quality of Television Pictures*, document Rec. ITU-R BT.500-14, International Telecommunication Union, Geneva, Switzerland, 2002. [Online]. Available: https://www.itu.int/dms_pubrec/itu-r/rec/bt/R-REC-BT.500-14-201910-I!!PDF-E.pdf

[40] S. Zielinski, F. Rumsey, and S. Bech, "On some biases encountered in modern audio quality listening tests-a review," *J. Audio Eng. Soc.*, vol. 56, no. 6, pp. 427–451, 2008.

[41] H. Z. Jahromi, D. T. Delaney, B. Rooney, and A. Hines, "Establishing waiting time thresholds in interactive Web mapping applications for network QoE management," in *Proc. 30th Irish Signals Syst. Conf. (ISSC)*, Jun. 2019, pp. 1–7.

[42] M.-J. Lobo, E. Pietriga, and C. Appert, "An evaluation of interactive map comparison techniques," in *Proc. 33rd Annu. ACM Conf. Hum. Factors Comput. Syst. (CHI)*, 2015, pp. 3573–3582.

[43] K. Hu, Z. Gui, X. Cheng, H. Wu, and S. McClure, "The concept and technologies of quality of geographic information service: Improving user experience of GIservices in a distributed computing environment," *ISPRS Int. J. Geo-Inf.*, vol. 8, no. 3, p. 118, 2019.

[44] B. Jenny, H. Jenny, and S. Räber, "Map design for the Internet," in *International Perspectives on Maps and the Internet* (Lecture Notes in Geoinformation and Cartography), M. P. Peterson, Ed. Berlin, Germany: Springer, 2008.

[45] T. Hoßfeld, M. Seufert, M. Hirth, T. Zinner, P. Tran-Gia, and R. Schatz, "Quantification of Youtube QoE via crowdsourcing," in *Proc. IEEE Int. Symp. Multimedia*, Dec. 2011, pp. 494–499.

[46] D. V. Cicchetti, "Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology," *Psychol. Assessment*, vol. 6, no. 4, p. 284, 1994.

[47] R. D. Portugal and B. F. Svaiter, "Weber-Fechner law and the optimality of the logarithmic scale," *Minds Mach.*, vol. 21, no. 1, pp. 73–81, Feb. 2011.

[48] S. Niida, S. Uemura, and H. Nakamura, "Mobile services," *IEEE Veh. Technol. Mag.*, vol. 5, no. 3, pp. 61–67, Sep. 2010.

[49] B. Shneiderman, "Response time and display rate in human performance with computers," *ACM Comput. Surv. (CSUR)*, vol. 16, no. 3, pp. 265–285, Sep. 1984.

HAMED Z. JAHROMI received the B.Sc. degree in software engineering technologies. He is currently pursuing the Ph.D. degree with the School of Computer Science, University College Dublin (UCD). He has over 12 years of industrial experience and worked as a Principal Network Engineer. He has been involved in many large scale networking projects. He is utilizing his industrial background to combine his extensive research in an enterprise setting, with an academic base at UCD. His current research interests include network QoE management in software defined networking (SDN), Internet traffic measurement, QoE/QoS, content-aware networking, and machine learning in communication networks.

DECLAN T. DELANEY received the Ph.D. degree in network analysis and design for the IoT from the School of Computer Science, UCD, in 2015. He previously worked at LMI Ericsson and collaborations with SMEs in H2020 funding proposals, where he maintains strong links with industry partners. He is currently an Assistant Professor at the School of Electrical and Electronic Engineering, UCD. He is also an SFI Funded Investigator on the project CONSUS, an SFI-industry funded collaboration focused on precision agriculture, and a Principal Investigator for the SmartBOG Project. His research interests are in the areas of network data analytics for adaptable programmable networks, and infrastructure for the IoT and sensor systems. He is currently involved in developing data assurance systems for sensed data and infrastructure design for collection and handling of multi streamed data for large precision agriculture. His other research interests include application aware networks and autonomous orchestration for 5G networks.

ANDREW HINES (Senior Member, IEEE) is currently the Director of the Research, Innovation and Impact for the School of Computer Science, UCD. He leads the QxLab Research Group with primary research interests in applying machine learning for applications in speech, audio, and video signal processing. He is a Funded Investigator in both the SFI CONNECT Research Centre for Future Networks and INSIGHT Research Centre for Data Analytics. He was a member of the management committee of the H2020 CryptoAction COST research network. He is a Committee Member of the Audio Engineering Society, Ireland. He has represented Ireland on the management committee of Qualinet, an FP7 European COST Action, where he led a task force focused on using machine learning to develop Quality of Experience (QoE) models.