Measuring Web Quality of Experience in Cellular Networks

Alemnew Sheferaw Asrese[1], Ermias Andargie Walelgne[1], Vaibhav Bajpai[2], Andra Lutu[4], Özgü Alay[3], and Jörg Ott[2]

[1] Aalto University
[2] Technische Universität München
[3] Simula Metropolitan
[4] Telefonica Research

Abstract. Measuring and understanding the end-user browsing Quality of Experience (QoE) is crucial to Mobile Network Operators (MNOs) to retain their customers and increase revenue. MNOs often use traffic traces to detect the bottlenecks and study their end-users experience. Recent studies show that Above The Fold (ATF) time better approximates the user browsing QoE compared to traditional metrics such as Page Load Time (PLT). This work focuses on developing a methodology to measure the web browsing QoE over operational Mobile Broadband (MBB) networks. We implemented a web performance measurement tool WebLAR (it stands for Web Latency And Rendering) that measures web Quality of Service (QoS) such as TCP connect time, and Time To First Byte (TTFB) and web QoE metrics including PLT and ATF time. We deployed WebLAR on 128 MONROE (a European-wide mobile measurement platform) nodes, and conducted two weeks long (May and July 2018) web measurement campaign towards eight websites from six operational MBB networks. The result shows that, in the median case, the TCP connect time and TTFB in Long Term Evolution (LTE) networks are, respectively, 160% and 30% longer than fixed-line networks. The DNS lookup time and TCP connect time of the websites varies significantly across MNOs. Most of the websites do not show a significant difference in PLT and ATF time across operators. However, Yahoo shows longer ATF time in Norwegian operators than that of the Swedish operators. Moreover, user mobility has a small impact on the ATF time of the websites. Furthermore, the website design should be taken into consideration when approximating the ATF time.

1 Introduction

Recent studies show that mobile data traffic is increasing exponentially, and web browsing is amongst the dominant applications on MBB networks [13]. The dependency on MBB networks and the widespread availability of LTE is boosting user expectations towards fast, reliable, and pervasive connectivity. The users make the MNOs responsible for the shortcomings in the mobile experience [5]. This demand pushes the MNOs to further enhance the capabilities of the mobile

networks for emerging applications. One of the challenging use cases for MBB networks is the mobility scenario [28], for example, browsing the web while commuting in a high-speed train. Thus, for MNOs, it is paramount to understand the end-user browsing experience while using their network [16]. Users are mostly concerned with the fulfillment of the quality expectation rather than the level of the QoS metrics like throughput.

There have been a number of previous efforts (§ 4) to measure and understand the performance of MBB networks. NetRadar [34, 37], SamKnows broadband measurement [12], Meteor [32] are some of the tools that have been developed to measure the QoS metrics from MBB network. These tools either aim at measuring the metrics related to QoS or do not indicate how the metrics are used to measure the QoE. Moreover, web performance and QoE have been well studied [3, 9, 13, 14, 19, 25–27, 33]. Nonetheless, most of the studies that investigated mobile web QoE are either from lab experiments or do not cover a wide range of metrics to approximate the end-user browsing experience. As a result, our understanding of web QoE on operational MNOs is limited. Mainly, this is because of two reasons: (1) the lack of large-scale measurements that investigate the application level metrics in operational MBB networks, and (2) the mapping of the network QoS to objective application QoS metrics and then to the subjective QoE, has not been well validated for mobile networks.

Our first contribution in this work (§ 2) is the design and development of *WebLAR* [7], a lightweight tool for measuring the end-user web experience over operational MNOs. The measurement tool can be deployed at scale and captures web latency and QoE metrics at different layers such as the DNS lookup time, TCP connect time, PLT, and the ATF time. The ATF time is the time required to show the content in the browsers' current viewport [15]. The authors in [9, 25] used two different approaches to approximate the ATF time in fixed-line networks. Asrese *et al.* [9] used a pixel-wise comparison of the changes in the browser's viewport to approximate the ATF time. They capture a series of screenshots of the webpage loading process and compare the pixel difference between consecutive screenshots with a three seconds threshold. When there is no change observed for three seconds, the webpage is considered as rendered completely. The ATF time is the difference between the starting time of the webpage loading process and the time where the last pixel change is observed. Hora *et al.* [25] used the browsers timing information to approximate the ATF time. They consider that the ATF time is the integral of the downloading time of the main HTML file, scripts, stylesheets and the images located in the above-the-fold area. By adopting the methods from the existing work [9, 25], we designed WebLAR to approximate the ATF time in operational MNOs. In addition, WebLAR captures network and device level metadata information such as the radio access technology, the GPS locations, CPU and memory usage in the device. Different confounding factors such as the device affect the QoE. In this work, we build a baseline view by using MONROE, a platform that can be used for performing measurements in a more controlled setting.

The second contribution of this work (§ 3) are the insights derived from the dataset collected using WebLAR . We deployed WebLAR on MONROE [6], a Europe-wide experimental platform for MBB network measurement. We measured the performance of eight popular websites from 128 stationary and mobile MONROE nodes distributed across Norway and Sweden. In our measurement campaign, measuring a larger set of websites was not possible because of data quota limitation. So, we picked eight websites (§ A) that are popular in Norway and Sweden. The result from our analysis shows that there is a difference in DNS lookup time, and TCP connect time of the websites across different MNOs. For most of the websites, there is no significant difference in PLT and ATF time across the operators. However, we also observed a big variation in ATF time of Yahoo between MNOs across different countries. That is, Yahoo has longer ATF time in the Norwegian MNOs. Moreover, we observed that user mobility does not have a significant effect on the web QoE.

The applicability of the aforementioned approaches [9,25] to approximate the ATF time have not been validated for webpages that have different design style. That is, one approach may work better for certain types of webpages but may not work well for others. Using the dataset collected using WebLAR, we showed that the website design should be taken into consideration while using the browser timing information and the pixel-wise comparison approaches to approximate the ATF time (§ 3.3). We also showed that for the pixel-wise comparison approach three seconds threshold is sufficient to determine when the content in the above-the-fold area of the webpage is stabilized. To encourage reproducibility [11], we open source the tool [7], and release the collected dataset along with the Jupyter notebooks [10] that were used for parsing and analysing the results.

2 Experiment Design

We begin by presenting our methodology (§ 2.1) to approximate the ATF time of websites. We provide details on the design, the experimental workflow (§ 2.2), and the implementation aspects (§ 2.3) of WebLAR required for its deployment on the MONROE platform.

2.1 Methodology

The contents in the *above-the-fold* area of the webpage (that is, the content within the current viewport of the browser) are the key parts of the webpage for the user to judge whether or not the page has downloaded and rendered. As such, the time at which the contents in the above-the-fold area stop changing and reach the final state is one objective metric to approximate the user QoE [15]. We refer to this as ATF time. One way to approximate the ATF time is by monitoring the pixel changes in the visible part of the webpage and detecting when it stabilizes [9]. Another method is approximating by using the performance timing information that the browsers provide [25]. Browsers provide APIs to retrieve performance and navigation time information of the websites.

The two approaches have their limitations. The webpage may not stabilize due to different reasons; for example, it may contain animating contents. As such, it might be difficult to detect when the webpage stabilizes. This makes it harder to approximate the ATF time using the pixel-wise approach. Conversely, in some cases it is difficult to identify the exact location of some types of objects. This is one of the challenges in approximating the ATF time using the browser's timing API. Thus, one approach could better approximate ATF time for certain types of websites, while the other approach may underestimate or overestimate it.

Recent studies [9,25] have developed tools to estimate the ATF time in fixed-line networks. We take this forward by designing and developing WebLAR that measures the web QoE in cellular networks by combining both approaches. WebLAR can approximate the ATF time using both the pixel-wise comparison [9] and using the browser performance timing information [25]. Unlike [9], where the measurement system approximates the ATF time by downloading all the web objects at the measurement nodes and pushing them to a centralized server location for processing, we approximate the ATF time at the MONROE nodes themselves. For simplicity of notations, we refer the ATF time approximated using this method as ATF_p time. Hora *et al.* [25] developed a Google Chrome extension to approximate the ATF time, which requires user interaction. Since the mobile version of Google Chrome does not support extensions (at least without using additional tools), it is not possible to use the browser timing information to approximate the ATF time in mobile devices. To close this gap, WebLAR approximates the ATF time in measurement probes that mimic mobile devices. We refer the ATF time approximated using this approach as ATF_b time. Moreover, using the browsers timing API, WebLAR also records metrics such as the DNS lookup time, TCP connect time, TTFB, and PLT. The browser API also enables us to get the web complexity metrics [22] including the number and the size of objects of the webpages. WebLAR also captures metadata information about the network conditions at the measurement nodes (e.g., MBB coverage profiles, signal strength) and other information that describe the user's mobility (e.g., GPS coordinates) and other events like CPU and memory usage.

2.2 Experiment workflow

Fig. 1 shows the sequence of operations of WebLAR experiment in MONROE measurement platform. The MONROE measurement platform provides a web interface where the users can submit their custom experiment (#1 in Figure). The MONROE back-end service then schedules (#2) the submitted user experiments to the selected nodes. It also starts the execution of the test according to the parameters that the user provided through the web interface. Once a node receives the commands for executing an experiment, it checks whether the `docker` container that contains the test is available locally. Otherwise, it fetches the `docker` container from a remote repository. Then the node starts the container with the parameters given in the MONROE web interface. When the container begins running the WebLAR experiment, WebLAR starts by checking the available network interfaces that have cellular connectivity and changes the default

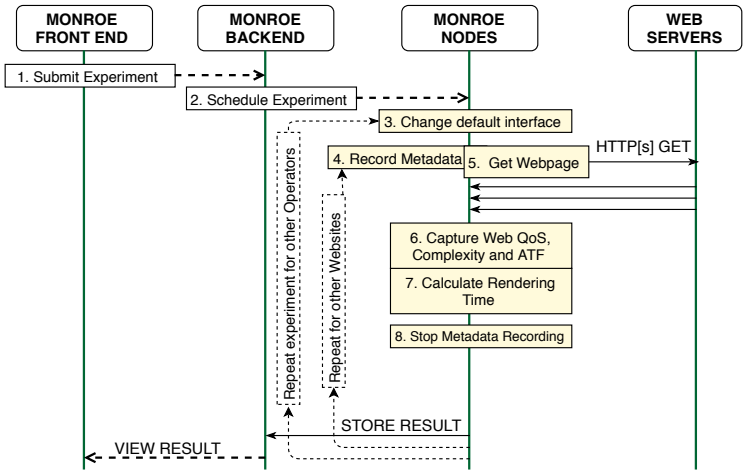

Fig. 1: Sequence diagram of the experiment using WebLAR tool in MONROE measurement platform.

gateway (#3) to one of the available interfaces to fetch the webpages. Then, the node immediately starts capturing the metadata information and simultaneously runs the Google Chrome browser (version 62) using Chromedriver (version 2.33) (#4 and #5). The Google Chrome browser starts in Incognito and maximized mode and with no-sandbox option. The browser issues `HTTP[S] GET` request to the given URL. When the browser starts downloading the webpage a video of the browsing session progress is captured for 30 seconds. Moreover, we capture the web QoS and complexity metrics of the webpage (#6) by using the browser timing information. At the same time, the ATF time is approximated using the timing information retrieved using the browser API. Once the browsing session is completed the recorded video is converted into a series of screenshots (bitmap images) in every 100 ms interval and the ATF time is calculated by comparing the pixel changes within the consecutive screenshots (#7). Then we stop capturing the metadata (#8) and send the results annotated with the metadata to the MONROE back-end. In one experiment submission, the steps from #3 to #8 may repeat depending on the number of cellular connectivity that the node has and the number of the webpages that the user wishes to measure. Finally, the user can retrieve the results from the MONROE back-end and can do analysis.

2.3 Implementation

The pixel-wise comparison approach: We designed a Java program that records a video (10 frames per second) of the browsing session on a predefined screen size. Then by using `ffmpeg` [23], the video is converted into bitmap images in 100 ms interval. `imagemagic` [1] is used to compare the pixel difference between consecutive images. Then we utilise a `python` script [9] to determine the ATF_p

time from the pixel differences. The ATF_p time is the point where there are no more pixel changes in consecutive X screenshots (*i.e.,* X/10 seconds threshold). A study [21] in 2016 shows the average PLT in 4G connection is 14 seconds. The study shows that more than half of the mobile users abandon the sites that take longer than three seconds to load. The study revealed that 75% of the mobile sites take longer than ten seconds to load. In the WebLAR experiment, we set three thresholds (3, 10 and 14 seconds) for declaring whether or not the webpage stabilizes. Hence, the ATF_p time is approximated with different webpage stabilizing thresholds.

Browser heuristic-based approach: We used the Google Chrome browser API and utilized the performance timing information to approximate ATF_b time using the browser's heuristic. First we detect all the resources of the website and their location on the webpage. Then, to approximate the ATF_b time, we integrate the download time of the images (that are located in the ATF area), javascript files, cascaded style sheet files, and the root document that contains the DOM structure of the webpage. Moreover, using the browser API, the QoS metrics such as the DNS lookup time, TCP connect time, TTFB, the DOM load time and PLT are captured. The web complexity metrics such as number and size of resources are also extracted using the API. We wrote a javascript implementation to approximate the ATF_b time and integrated it within the Java program used to approximate the ATF_p time.

3 Analysis

We begin by presenting the dataset (§ 3.1) we collected after deploying WebLAR on the MONROE platform. We present the analysis using this dataset, focussing on IP path lengths (§ 3.2), web latency and QoE (§ 3.3) and specifically QoE under mobility (§ 3.4) conditions.

3.1 Dataset

We ran the WebLAR experiment for two weeks (May 19 - 26, 2018 and July 2 - 9, 2018) in 128 MONROE nodes located in Norway and Sweden. The nodes are equipped with one or two SIM cards with 4G connectivity. Nine of the nodes deployed in Norway are connected with a Swedish operator roaming [29] in Norway. Our measurement campaign covers a total of six operators. During the campaign, nodes are set to fetch specific pages of eight popular websites (A). The WebLAR experiment execute every six hours. In the rest of this paper, we refer to the websites with the name of their base URL. We performed pre-processing to prune out results where the experiment failed to report values of all metrics (e.g., due to browser timeout settings) leaving us with ∼18K data points.

3.2 IP path lengths

We began by analysing the IP paths towards the measured websites. WebLAR uses `traceroute` to measure the IP path length and the round trip time towards the websites. To study the IP path length and the latency difference in LTE and

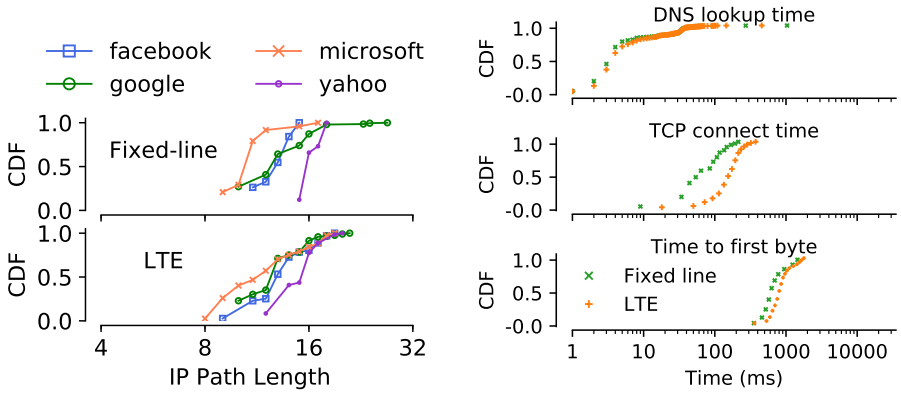

Fig. 2: The distribution of (1) IP path length and (2) web QoS metrics from fixed-line and LTE broadband networks as observed from selected 29 nodes.

fixed-line networks, we ran WebLAR on 29 MONROE nodes in Italy, Norway, Spain, and Sweden. Fig. 2 (1) shows the IP path length towards selected websites in fixed-line and LTE networks from 29 MONROE nodes. The result shows that in the median case, the IP path length in LTE and fixed-line network is similar.

3.3 Web latency and QoE

Fig. 2 (2) shows the latency towards the websites from fixed-line and LTE networks from 29 MONROE nodes. We observe that there is no significant difference in the DNS lookup time and PLT (not shown) of the websites from fixed-line and LTE network. However, the TCP connect time and TTFB of the websites are shorter in fixed-line network. For instance, in the median case, in LTE network the TCP connect time, and TTFB are respectively, 160% and 30% longer than that observed in fixed-line networks. Due to security reason, the browser timing API gives the same value for the start and end of the TCP connect and DNS lookup time for cross-origin resources. That is, unless the user explicitly allows the server to share these values, by default the TCP connect time and DNS lookup time is 0 for the cross-origin resources [30]. As a result, three websites (Google, Microsoft, and Yahoo) report 0 for these metrics. The discussion of the DNS lookup time and TCP connect time does not include these three websites.

Fig. 3 (1) shows the latency of the websites under different MNOs. Note, the Norwegian and Swedish operators are labeled with NO_o and SE_o, respectively, where $o \in \{1, 2, 3\}$. SE_r refers to a Swedish operator roaming in Norway. The result shows the MNOs have different performance in terms of DNS lookup time (ranges from 35ms to 60ms, in the median case) and TCP connect time (ranges from 100 ms to 200ms, in the median). One of the causes for the variation in the DNS lookup time across the MNOs could be attributed to the presence of cached DNS entries [36]. The result also shows that, the difference in TTFB and PLT

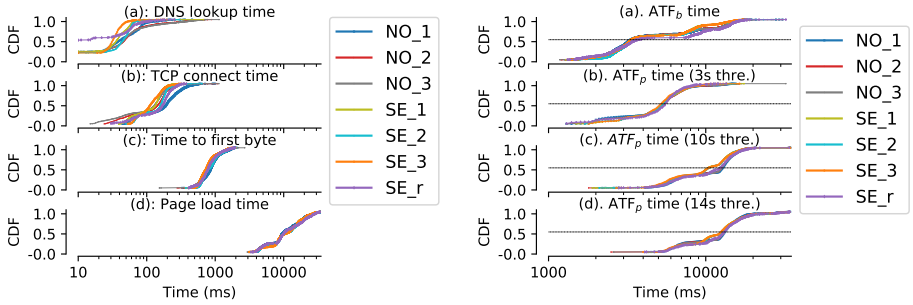

Fig. 3: The distribution of (1) DNS lookup time, TCP connect time, TTFB, and PLT and (2) ATF time as approximated using the two approaches.

of the websites across different MNOs is not high (*i.e.,* in the median case, only 200ms to 600 ms difference in PLT). We applied Kolmogorov - Smirnov test to investigate the significance of the difference in PLT across MNOs. In most of the cases, we found a smaller p-value (below 0.05) between the PLT of the websites across MNOs. This confirms that there is a difference in PLT of the websites across MNOs. We also found a higher p-value between PLT across MNOs within the same country (*e.g.,* 0.46 between NO_2 and NO_2, 0.4 between SE_1 and SE_3). This observation indicates that MNOs within the same country exhibit similar PLT towards these websites. The result also shows that there is up to 1 second improvement in the PLT compared with a previous [21] observations.

Fig. 3 (2) shows the distribution of the ATF time towards websites across different MNOs as approximated using the two approaches. Fig. 3 (2, top) shows the approximated ATF_b time. The long tails of the distribution in this result is due to Facebook and BBC, which have higher number of objects and overlapping images in the above-the-fold area. Fig. 3 (2, bottom 3) show the ATF_p with three, ten and 14 seconds threshold, respectively. From the result, we can see that in the median case, the ATF_b is shorter than the ATF_p time with three seconds threshold. This indicates that three seconds is a sufficient threshold to declare whether the website has stabilized or not. As such, going forward, we only consider three seconds threshold for approximating the ATF time using the pixel-wise comparison approach. The difference in the ATF time of the websites across most of the MNOs is small (*i.e.,* in the median case, the difference is 100 ms to 300ms). However, we notice that the difference in ATF time between SE_r and the other MNOs is large (*i.e.,* in the median case, ATF_b time can be up to 400 ms and ATF_p time can be up to 4200 ms). By applying a Kolmogorov - Smirnov test, we found a smaller p-value (below 0.05) between the ATF_b time of the different MNOs. This confirms that there is a difference between ATF_b times across MNOs. Only the ATF_b time of websites between SE_1 and SE_3 shows a p-value of 0.75, highlighting similar QoE between the two MNOs.

We also analysed the rendering performance of each website. Fig. 4 (1) shows the distribution of the ATF time approximated using the two approaches and the PLT of the websites. Through manual inspection, we observed that some of the

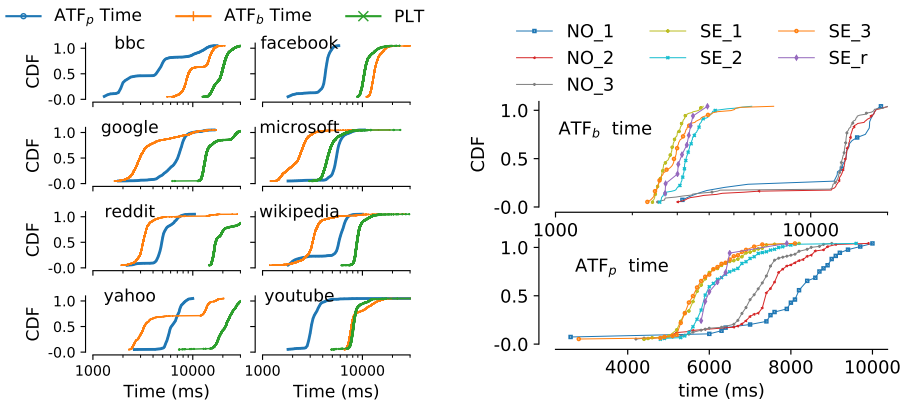

Fig. 4: (1)The CDF of the PLT and the ATF time of the different websites. (2) The ATF time of Yahoo across different MNOs.

websites, e.g., Microsoft, have a fewer number of objects and take shorter time to show the contents of the above-the-fold area. The ATF approximation using both approaches confirms this. On the contrary, websites like Facebook have multiple objects located in the above-the-fold area (confirmed through manual inspection). The objects may overlap each other where some of the objects may not be visible in the front unless the user takes further action (e.g., clicking the sliding button). In such cases, the browser heuristic based ATF time approximation overestimates the ATF time. Hence, for these kinds of websites, the ATF time approximation based on the browser heuristic does not better represent the end user experience. That is, the missing or delay in the download of those overlapped objects do not have effect in the visual change of the websites. Therefore, for the websites that have overlapping objects in the above-the-fold area, the ATF time needs to be approximated in a different way. For instance, Fig. 4 (1) shows that the ATF_p time of Facebook is below half of its PLT, which is much shorter than its ATF_b time. This shows that the pixel-wise comparison approach of ATF time approximation is better for websites that have overlapping contents. However, approximating the ATF time using the pixel-wise comparison approach may also overestimate the ATF time for some websites. For instance, Microsoft has fewer images in the above-the-fold area, and the ATF_b time is short. However, the visual look of the webpage seems to be manipulated by using css and javascripts and have animating contents. As a result, the pixel-wise comparison approach yields longer ATF time for this website. Therefore, the design of the website can have an impact on the two ATF time approximation methods. Furthermore, due to the design pattern adopted by some websites, the objects are fetched asynchronously and the TCP connection may not be closed. As such, the javascript `onLoad` event may fire before all the objects are fetched. In such cases, the ATF_b time is longer than that of the PLT.

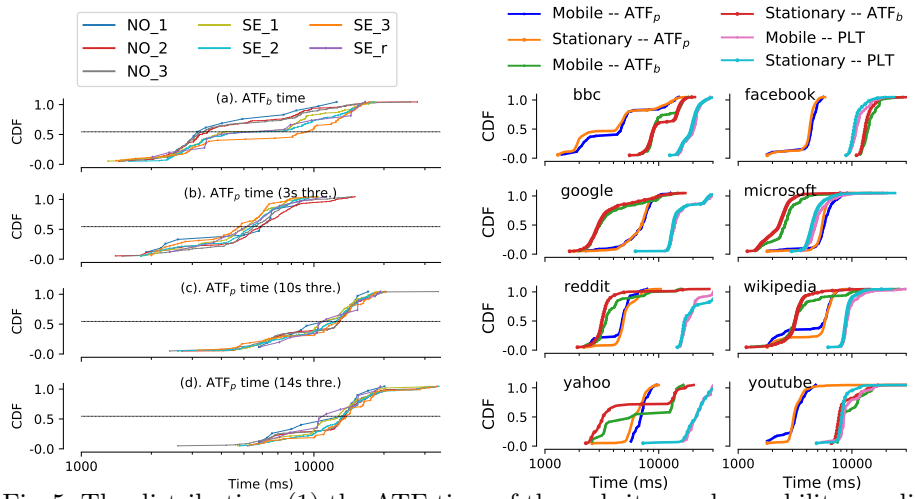

Fig. 5: The distribution: (1) the ATF time of the websites under mobility condition across different operators, and (2) The ATF time and PLT of the websites under different mobility conditions.

Fig. 4 (1) also shows that the ATF time of BBC, Yahoo and Wikipedia exhibits a bimodal distribution. We investigated this aspect further by observing the ATF time of these websites from different operators. For instance, Fig. 4 (2) shows the distribution of the ATF time of Yahoo across the different MNOs approximated using the two approaches. The result reveals that in the Norwegian MNOs, Yahoo takes longer to show the contents in the above-the-fold area. As such, the bimodal distribution of ATF time is due to the difference observed in the operators across different country. The impact of the longer download time of the objects in the above-the-fold area is reflected in the ATF_p time of the websites. For the other two websites we see a difference across the operators. That is, the bimodal distribution happens in all operators. Fig. 4 (2) and Fig. 3 (1) also show that the Swedish operator roaming in Norway has a similar QoE with the native Swedish operator. As such, the home-routed roaming [29] configuration does not have much impact on the QoE when the user travels relatively small distances (*i.e.,* between Norway and Sweden).

3.4 Web QoE under mobility conditions

Fig. 5 (1) shows the distribution of the ATF time of the websites under mobility scenario as approximated using the two methods. The results show that ATF time of the websites measured from nodes deployed in trains and buses are similar to that of the nodes deployed in homes and offices. However, the variation in ATF time across different MNOs is relatively higher under mobility scenario.

The nodes deployed in trains can be online even though the trains are at the garage; hence some nodes may not be moving in some cases. Fig. 5 (2) shows the ATF time and PLT of websites from buses and trains which were moving while

the measurement was conducted. The result shows that most of the websites have almost similar PLT in a mobile and a stationary situation. However, the ATF time of some of the websites is relatively longer in mobility scenario. For instance, in the median case, the ATF time of Microsoft, Yahoo, Reddit, and Facebook is 0.3 to 1 second longer under mobility condition. Yahoo shows different behavior in the ATF time from stationary and mobile nodes. That is, 60% of the measurements from the mobiles nodes, and 40% of the measurements from the stationary nodes show a drastic change (more than 7 seconds difference) of the ATF time. To understand the causes for this drastic change we analyzed the ATF time of this website at each operator. We found that in the Norwegian operators Yahoo takes longer time to show the contents in the above-the-fold area. One of the causes for this could be the IP path length between the operators and the Yahoo content server. Using a `traceroute` measurement we analyzed the IP path lengths that the nodes traverse to reach the web servers from different locations. We observed that the nodes hosted in Norwegian operators traverse up to 20 IP hops to reach the Yahoo web server. Instead, other Swedish operators take a maximum of 16 IP hopes to reach Yahoo's web server.

4 Related Work

The web has been well studied. Various web QoE measurement tools and methodologies are available [8, 9, 25, 35]. Most of these tools focus on fixed-line networks. For instance, Varvello *et al.* [35] designed *eyeorg*, a platform for crowdsourcing web QoE measurements. The platform shows a video of the page loading progress to provide a consistent view to all the participants regardless of their network connections and device configurations. Unlike *eyeorg*, our measurement tool does not require user interaction to evaluate the web QoE, rather it uses different approaches to approximate the web QoE. Cechet *et al.* [18] designed mBenchLab that measure web QoE in smartphones and tablets by accessing cloud hosted web service. They measured the performance of few popular websites and identify the QoE issues observing the PLT, the traditional web QoE metric. Casas *et al.* [17] studied the QoE provisioning of popular mobile applications using subjective laboratory tests with end-device through passive measurement. They also studied QoE from feedback obtained in operational MNOs using crowd-sourcing. They showed the impact of access bandwidth and latency on QoE of different services including web browsing on Google Chrome.

Balachandran *et al.* [13] proposed a machine learning approach to infer the web QoE metrics from the network traces, and studied the impact of network characteristics on the web QoE. They showed that the web QoE is more sensitive for the inter-radio technology handover. Improving the signal to noise ratio, decreasing the load and the handover can improve the QoE. Ahmad *et al.* [4] analyzed call-detail records and studied WAP support for popular websites in developing regions. Nejati *et al.* [31] built a testbed that allows comparing the low-level page load activities in mobile and non-mobile browsers. They showed that computational activities are the main bottlenecks for mobile browsers, which indicates that browser optimizations are necessary to improve the mobile web

QoE. Dasari *et al.* [20] studied the impact of device performance on mobile Internet QoE. Their study revealed that web applications are more sensitive for low-end hardware devices compared to video applications.

Meteor [32] is a measurement tool which determines the speed of the network and estimates the experience that the user can expect while using selected popular applications given their connection requirements. The methodology used by Meteor is not open aside from the high-level explanation of the system. It is not clear how the expected experience is computed and which performance metrics are used for a given application. Perhaps, it is based on QoS metrics like throughput and latency test, which may not be the only factors that affect the performance of different application [20]. Unlike Meteor, we measure different metrics at the network and application level, e.g., TTFB, PLT, as well as ATF time at the browser which is more important from the user perspective. WebPageTest [2] and Google Lighthouse [24] are other tools designed to assess the web performance from different locations using different network and device types. These tools measure PLT, SpeedIndex, TTFB, time to visually complete (TTVC), first contentful paint (FCP), first meaningful paint (FMP), time to interactive (TTI), and last visual change metrics. WebLAR measures the ATF time, but it does not measure SpeedIndex, TTVC, TTI, and FCP yet. SpeedIndex [3] is a metric proposed by Google to measure the visual completeness of a webpage. It can be approximated either by capturing video of the webpage download progress or by using the paint events exposed by Webkit. We make WebLAR publicly available [7] and invite the measurement community for contributions to help improve this tool.

5 Conclusions

We presented the design and implementation of WebLAR – a measurement tool that measures web latency and QoE in the cellular network. We applied ATF time as the metric to to approximate the end-user experience. We followed two different approaches to approximate the ATF time: pixel-wise comparison and the browser heuristics. We deployed WebLAR on the MONROE platform for two weeks. The results show that the DNS lookup time and PLT of the selected websites have similar performance in LTE and fixed-line networks. However, the TCP connect time and TTFB of the websites are longer in LTE networks. Moreover, the DNS lookup time and TCP connect time of the websites varies across MNOs. For most of the websites, PLT, and ATF time do not have a significant difference across operators. We observed that mobility has small impact on the ATF time of the websites. We also showed that the design of the website should be taken into account when using two approaches to approximate the ATF time.

Limitations and Future Work: We only measured eight websites in this study and did not perform a subjective QoE evaluation. We also did not consider the impact of device capabilities on the web QoE since our measurement nodes were homogenous. In the future, we plan to extend WebLAR to capture other metrics such as RUM SpeedIndex, TTI, first contentful paint and also evaluate the ATF time using different screen sizes.

References

1. ImageMagick: Tool to create, edit, compose, or convert bitmap images. https://imagemagick.org, Retrieved October 12, 2018
2. WebPageTest. https://www.webpagetest.org, Retrieved on Jan 09, 2019
3. WebPagetest Metrics: SpeedIndex. https://sites.google.com/a/webpagetest.org/docs/using-webpagetest/metrics/speed-index, Retrieved on Oct 15, 2018
4. Ahmad, S., Haamid, A.L., Qazi, Z.A., Zhou, Z., Benson, T., Qazi, I.A.: A View from the Other Side: Understanding Mobile Phone Characteristics in the Developing World. ACM IMC (2016), http://dl.acm.org/citation.cfm?id=2987470
5. Akamai White Paper: Measuring Real Customer Experiences over Mobile Networks. https://www.akamai.com/jp/ja/multimedia/documents/white-paper/measuring-real-customer-experiences-over-mobile-networks-report.pdf, [Retrieved: Oct 12, 2017]
6. Alay, Ö., Lutu, A., Quirós, M.P., Mancuso, V., Hirsch, T., Evensen, K., Hansen, A.F., Alfredsson, S., Karlsson, J., Brunstrom, A., Khatouni, A.S., Mellia, M., Marsan, M.A.: Experience: An Open Platform for Experimentation with Commercial Mobile Broadband Networks. ACM MobiCom (2017), http://doi.acm.org/10.1145/3117811.3117812
7. Asrese, A.S.: WebLAR: A Web Performance Measurment Tool. https://github.com/alemnew/weblar (2019)
8. Asrese, A.S., Eravuchira, S.J., Bajpai, V., Sarolahti, P., Ott, J.: Measuring Web Latency and Rendering Performance: Method, Tools & Longitudinal Dataset. IEEE Transactions on Network and Service Management (2019, to appear)
9. Asrese, A.S., Sarolahti, P., Boye, M., Ott, J.: WePR: A Tool for Automated Web Performance Measurement. IEEE Globecom Workshops (2016), https://doi.org/10.1109/GLOCOMW.2016.7849082
10. Asrese, A.S., Walelgne, E., Bajpai, V., Lutu, A., Alay, Ö., Ott, J.: Measuring Web Quality of Experience in Cellular Networks (Dataset). https://github.com/alemnew/2019-pam-weblar (2019)
11. Bajpai, V., Kühlewind, M., Ott, J., Schönwälder, J., Sperotto, A., Trammell, B.: Challenges with reproducibility. pp. 1–4. SIGCOMM Reproducibility Workshop (2017), https://doi.org/10.1145/3097766.3097767
12. Bajpai, V., Schönwälder, J.: A Survey on Internet Performance Measurement Platforms and Related Standardization Efforts. IEEE Communications Surveys and Tutorials **17**(3) (2015), https://doi.org/10.1109/COMST.2015.2418435
13. Balachandran, A., Aggarwal, V., Halepovic, E., Pang, J., Seshan, S., Venkataraman, S., Yan, H.: Modeling Web Quality-of-Experience on Cellular Networks. ACM MobiCom (2014), http://doi.acm.org/10.1145/2639108.2639137
14. Barakovic, S., Skorin-Kapov, L.: Multidimensional Modelling of Quality of Experience for Mobile Web Browsing. Computers in Human Behavior **50** (2015), https://doi.org/10.1016/j.chb.2015.03.071
15. Brutlag, J., Abrams, Z., Meenan, P.: Above the Fold Time: Measuring Web Page Performance Visually. https://conferences.oreilly.com/velocity/velocity-mar2011/public/schedule/detail/18692
16. Cao, Y., Nejati, J., Wajahat, M., Balasubramanian, A., Gandhi, A.: Deconstructing the Energy Consumption of the Mobile Page Load. ACM Proceedings of the ACM on Measurement and Analysis of Computing Systems **1**(1) (2017), http://doi.acm.org/10.1145/3084443

17. Casas, P., Seufert, M., Wamser, F., Gardlo, B., Sackl, A., Schatz, R.: Next to You: Monitoring Quality of Experience in Cellular Networks From the End-Devices. IEEE Transactions on Network and Service Management **13**(2) (2016), https://doi.org/10.1109/TNSM.2016.2537645
18. Cecchet, E., Sims, R., He, X., Shenoy, P.J.: mBenchLab: Measuring QoE of Web Applications using Mobile Devices. International Symposium on Quality of Service (IWQoS) (2013), https://doi.org/10.1109/IWQoS.2013.6550259
19. Chen, Q.A., Luo, H., Rosen, S., Mao, Z.M., Iyer, K., Hui, J., Sontineni, K., Lau, K.: QoE Doctor: Diagnosing Mobile App QoE with Automated UI Control and Cross-layer Analysis. ACM Internet Measurement Conference (2014), http://doi.acm.org/10.1145/2663716.2663726
20. Dasari, M., Vargas, S., Bhattacharya, A., Balasubramanian, A., Das, S.R., Ferdman, M.: Impact of Device Performance on Mobile Internet QoE. pp. 1–7. Internet Measurement Conference (2018), https://doi.org/10.1145/3278532.3278533
21. DoubleClick: The Need for Mobile Speed: Better User Experiences, Greater Publisher Revenue. https://goo.gl/R4Lmfh, Retrieved Feb 26, 2018
22. Eravuchira, S.J., Bajpai, V., Schönwälder, J., Crawford, S.: Measuring Web Similarity from Dual-stacked Hosts. pp. 181–187. Conference on Network and Service Management (2016), https://doi.org/10.1109/CNSM.2016.7818415
23. FFmpeg: FFmpeg: A complete, cross-platform solution to record, convert and stream audio and video. https://ffmpeg.org, Retrieved October 12, 2018
24. Google: Lighthouse: An open-source, automated tool for improving the quality of web pages. https://developers.google.com/web/tools/lighthouse, Retrieved on Jan 09, 2019
25. da Hora, D.N., Asrese, A.S., Christophides, V., Teixeira, R., Rossi, D.: Narrowing the Gap Between QoS Metrics and Web QoE Using Above-the-fold Metrics. Passive and Active Measurement (2018), https://doi.org/10.1007/978-3-319-76481-8_3
26. Hosek, J., Ries, M., Vajsar, P., Nagy, L., Sulc, Z., Hais, P., Penizek, R.: Mobile web QoE study for smartphones. IEEE GLOBECOM Workshop (2013), https://doi.org/10.1109/GLOCOMW.2013.6825149
27. Hoßfeld, T., Metzger, F., Rossi, D.: Speed Index: Relating the Industrial Standard for User Perceived Web Performance to web QoE. IEEE International Conference on Quality of Multimedia Experience (2018), https://doi.org/10.1109/QoMEX.2018.8463430
28. Li, L., Xu, K., Wang, D., Peng, C., Zheng, K., Mijumbi, R., Xiao, Q.: A Longitudinal Measurement Study of TCP Performance and Behavior in 3G/4G Networks Over High Speed Rails. IEEE/ACM Transactions on Networking **25**(4) (2017), https://doi.org/10.1109/TNET.2017.2689824
29. Mandalari, A.M., Lutu, A., Custura, A., Khatouni, A.S., Alay, Ö., Bagnulo, M., Bajpai, V., Brunström, A., Ott, J., Mellia, M., Fairhurst, G.: Experience: Implications of Roaming in Europe. pp. 179–189. MOBICOM (2018), https://doi.org/10.1145/3241539.3241577
30. Mozila: Using the Resource Timing API. https://developer.mozilla.org/en-US/docs/Web/API/Resource_Timing_API/Using_the_Resource_Timing_API, Retrieved May 24, 2018
31. Nejati, J., Balasubramanian, A.: An In-depth Study of Mobile Browser Performance. In: Conference on World Wide Web. pp. 1305–1315 (2016), https://doi.org/10.1145/2872427.2883014
32. OpenSignal: Meteor. https://meteor.opensignal.com, [Retrieved: May 12, 2017]

33. Sackl, A., Casas, P., Schatz, R., Janowski, L., Irmer, R.: Quantifying the impact of network bandwidth fluctuations and outages on Web QoE. IEEE International Workshop on Quality of Multimedia Experience (2015), https://doi.org/10.1109/QoMEX.2015.7148078
34. Sonntag, S., Manner, J., Schulte, L.: Netradar - Measuring the Wireless World. IEEE International Symposium and Workshops on Modeling and Optimization in Mobile, Ad Hoc and Wireless Networks (2013), http://ieeexplore.ieee.org/document/6576402/
35. Varvello, M., Blackburn, J., Naylor, D., Papagiannaki, K.: EYEORG: A Platform For Crowdsourcing Web Quality Of Experience Measurements. ACM Conference on emerging Networking EXperiments and Technologies (2016), http://doi.acm.org/10.1145/2999572.2999590
36. Walelgne, E.A., Kim, S., Bajpai, V., Neumeier, S., Manner, J., Ott, J.: Factors Affecting Performance of Web Flows in Cellular Networks. IFIP Networking (2018)
37. Walelgne, E.A., Manner, J., Bajpai, V., Ott, J.: Analyzing Throughput and Stability in Cellular Networks. pp. 1–9. IEEE/IFIP Network Operations and Management Symposium (2018), https://doi.org/10.1109/NOMS.2018.8406261

Appendix A List and category of measured webpages

The websites are selected from different categories such as social media, news websites, and WIKI pages. Moreover, while selecting these websites, the design of the websites (from simple to media-rich complex webpages) and the purpose of the websites are taken into consideration. Furthermore, for each website we selected a specific webpage that does not require user interaction to show meaningful contents to the user.

- News websites
  - http://www.bbc.com
  - https://news.google.com
- Wiki websites
  - https://en.wikipedia.org/wiki/Alan_Turing
  - https://www.reddit.com
- Social media websites
  - https://www.youtube.com
  - https://www.facebook.com/places/Things-to-do-in-Paris-France/110774245616525
- General websites
  - https://www.microsoft.com
  - https://www.yahoo.com

Appendix B Additional Observations

Although not specific to mobility scenario, Fig. 5 (2) also shows that PLT can under– or over–estimate the web QoE. For instance, for Facebook, the onLoad event fires before all the necessary web objects in the above-the-fold area are downloaded. For these types of websites the PLT underestimates the user QoE. On the other hand, for websites like Yahoo and Reddit, the ATF is shorter compared with PLT time, which overestimates the user QoE.