Back to Normal? Impact of Temporally Increasing Network Disturbances on QoE

Junaid Shaikh, Markus Fiedler, Pangkaj Paul
Blekinge Institute of Technology
Karlskrona, Sweden
Email: {junaid.junaid, markus.fiedler}@bth.se,
papb11@student.bth.se

Sebastian Egger
Telecom Research Center
FTW, Vienna, Austria
Email: egger@ftw.at

Frederic Guyard
R&D Orange Labs
Sophia Antipolis, France
Email: frederic.guyard@orange.com

Abstract—Brief episodes of network faults and performance issues adversely affect the user Quality of Experience (QoE). Besides damaging the current opinions of users, these events may also shape user's future perception of the service. Therefore, it is important to quantify the impact of such events on QoE over time. In this paper, we present our findings on the temporal aspects of user feedback to disturbances on networks. These findings are based on subjective user tests performed in the context of web browsing on an e-commerce website. The results of this study suggest that the QoE drops significantly every time the page load time grows. The after-effects of network disturbances on user QoE remain visible even when the network problems are over, i.e., users do not immediately return to the same level of opinion scores as compared to the corresponding pre-disturbance phase. They tend to remember their recent experiences. Our results also show that there are four segments of users that exist with regards to their feedback to page load times. Network operators may customize their services according to each segment of users to raise the overall QoE. Finally, we show that the exponential relationship provides best fits of QoE and page load times for all segments of users.

Keywords—*IP Network, Outages, OFF times, Temporal QoE, Quality of Experience, Quality of Service, Memory effect, Web browsing*

I. INTRODUCTION

Increasing reliance of a wide spectrum of daily life activities on the Internet put stringent requirements on today's data networks. They need not only be available and accessible around the clock but also capable enough in delivering distinct quality. However, unfortunately, the networks still are not vigilant enough to meet these demands. Particularly, the fast emerging mobile broadband networks are prone to failures and transient outages, mainly due to resource allocation, mobility and configuration issues, which debase Quality of Experience (QoE).

Generally network downtimes over large time windows (minutes to hours scales) are noticed and resolved well by service providers. However, the issues related to transient outages often remain unnoticed [1][2]. Eventually, the ON and OFF phases – giving rise to delays and waiting times for users – without any perceptible follow-up by service providers, often result in dissatisfied users. Thus, there are studies needed to understand the accumulating impact of recurring short-term service disruptions (on the scale of seconds) on temporal QoE.

Previously, a number of studies has been done to quantify the impact of delays on QoE [3]–[5]. Similarly, relationships between QoE and Quality of Service (QoS) are also derived [6]–[9]. Yet, there is a lack of studies, which quantify the accumulating impact of service disruptions on QoE over time. In another paper [10], authors have presented results with regards to the time dynamics of QoE. Their investigation showed the role of user memory in user perception of waiting times on the Web. However, there are further studies needed in this regard to portray the escalating effect of short-term outages on user QoE.

In this paper, we have tried to illustrate the piling-up effect of bad memories from increasing network OFF times on the user QoE. Our study is based on task-based user subjective tests done in the context of web browsing on an e-commerce website. This study attempts to answer the possible underlying psychological factors, which motivate users to adopt a certain opinion in case of delays. One of the important aspects related to this is the assessment of user's tendency to return to the pre-disturbance level of opinion about service after network problems are rectified. Using clustering, we also show different segments of users in terms of their memory and response to the delays on web pages. The knowledge about such aspects can be instrumental for service providers in customizing their services according to the user types and, hence, in devising better strategies for retaining their customers.

The structure of this paper is as follows. Section II presents the methodology of user tests. Section III provides details about the experiment setup used in this study to conduct user tests. Section IV presents results on the users' responses to Page Load Times (PLTs) and the possible reasoning behind the obtained results. Finally, Section V poses a set of conclusions and an outlook of the future work.

II. METHODOLOGY

In this study, a total of 43 subjects participated in subjective tests. The mean age of participants was 24.5 years, the maximum age was 32 years and minimum age was 21 years. All the subjects were everyday users of the web browsing service and use e-commerce websites regularly for online shopping. Before starting the test, a 5-minute training session was conducted for each of the participants, in which necessary instructions were provided about the test procedure.

Each subject in this study performed task-based web browsing for an e-commerce website. The task was based on

the selection and purchase of a laptop computer. Each subject went through 12 shopping sessions. Each shopping session was based on web browsing of three web pages: Product selection page, product details & purchase page and payment confirmation page. The first page (product selection) consisted of 21 objects (2 CSS, 2 JS, 17 JPEG and PNG images). The second page (product details & purchase) consisted of 3 objects (all images) and some text providing the specification of selected product. The third webpage of shopping session consisted of 2 objects (all images) and some text acknowledging the purchase.
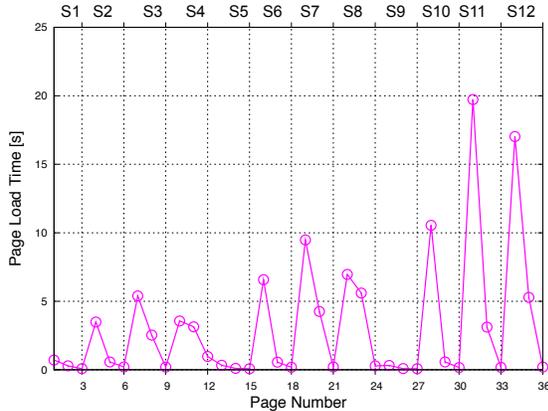

Fig. 1.   Page Load Time (PLT) per web page

Particular packets carrying web page content on the network were targeted and delayed in order to increase Page Load Times (PLTs). These packet-based delays introduced OFF times resembling outages on real network [1][2]. Figure 1 illustrates the PLTs over time faced by each subject. The x-axis of the plot in Figure 1 represents web page number and the y-axis represents the PLTs. Each shopping session (based on three web pages) is illustrated by S1 to S12 on the secondary x-axis. As illustrated in Figure 1, each user went through a variety of PLTs, from less than 1 s to around 20 s. Delays were introduced in both increasing as well as decreasing order to understand the transition in opinion scores of subjects. For instance, the PLT was increased at page 4 and then decreased at page 5 to assess how subjects react to them.

At every page, each subject was asked the following two questions:

1) *Which web page is this?*

    The following three options were given to answer the above question:
    - *Product Selection*
    - *Product Details & Purchase*
    - *Purchase Confirmation*

    This question to test whether subjects are attentive to the their task.
2) *How do you feel about its loading time?*
    The answer of this question was provided in the

form of the five-level ACR scale for rating quality, recommended by ITU-T.
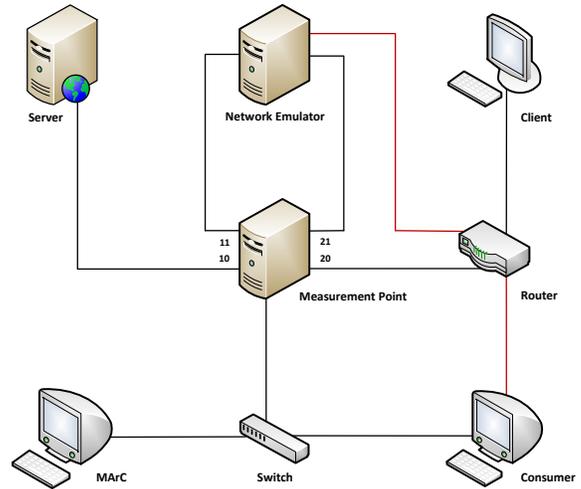
III.   EXPERIMENTAL SETUP

Fig. 2.   Experiment setup

A client-server model was implemented to conduct experiments. When a subject requested for a web page from client machine, the server received the request and subsequently, responded with content of the requested web page. These requests and responses were transferred via a network emulator called KauNet that allows for having impact on specific packets, thus controlling the PLTs in well-defined ways [11].

The server was installed with the Ubuntu 10.10 operating system. It was configured with Apache 2.2 to act as a web server. Apache is currently the most popular web server [12]. Web pages deployed on the web server were developed using CodeIngter (a PHP framework) [13]. Moreover, the widely used free open-source software for Linux systems, Bind9 was installed to setup the DNS server in order to translate user-requested URL to web server IP address.

Similarly, the client machine consisted of Windows 7 operating system. The Google Chrome web browser was installed and used for web browsing on the client side. Google Chrome was chosen because currently it is far more popular than any other web browser [15]. Moreover, an open-source web debugging proxy tool called Fiddler [16] was also deployed on the client side in order to collect the HTTP (S) logs. These logs were collected in Java Script Object Notation (JSON) format and stored in HTTP ARchive (HAR) files.

In order to collect and store the network level traffic, the Distributed Passive Measurement Infrastructure (DPMI) was deployed [14]. As shown in the Figure 2, the DPMI consisted of Measurement Area Controller (MArC), Measurement Points (MPs) and Consumer machines to control, capture and store the network traffic, respectively. The MPs were equipped with two Endace Data Acquisition and Generation (DAG) 3.5E cards to capture network traffic near the server and the client sides in both directions. The choice of DPMI is motivated
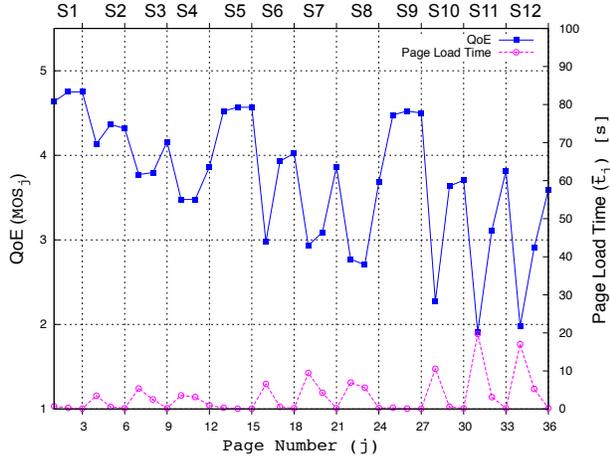
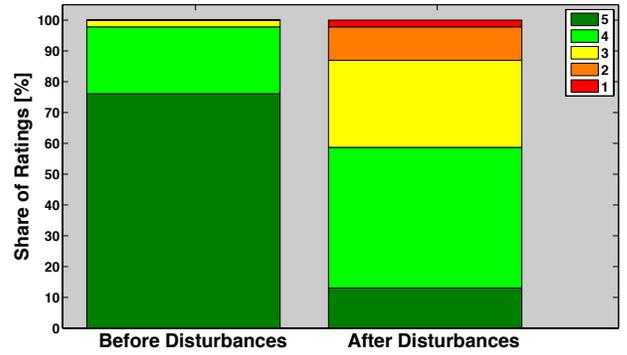Fig. 3. Mean Opinion score (MOS) and mean PLTs per web page

Fig. 4. Opinion scores for undisturbed web page transfers (Left: Rating share for page 1 [Mean PLT: 0.7 s], Right: Rating share for page 36 [Mean PLT: 0.2 s])

by the fact that it enables high accuracy measurements (up to nanosecond level) with a distributed architecture to collect packets at multiple points within a network.

A signaling script was developed and placed on the client machine in order to make the test procedure automatic. Based on the design of experiment procedure (Session ID and URL), this script signaled the desired network settings to the network emulator. Similarly, it signaled other machines on the network to collect required logs in appropriate files based on the User ID, Session ID and the URL. Furthermore, it collected the answers to the questions, mentioned in the previous section and stored them in a local data base along with web page URLs, network settings and User IDs. The automatic setup has proved helpful in preventing the interruptions for subjects during the tests that would have otherwise occurred because of manually changing network settings, creating log files and collecting opinion scores from subjects.

IV. RESULTS AND ANALYSIS

A. QoE over time

Each subject browsed through a total of 36 web pages and provided a rating on the MOS scale ranging from 5 (= excellent) to 1 (=bad) after a web page was completely loaded. A total of 44 ratings and PLTs per web page were obtained from the 44 participants. Subsequently, the means of opinion scores (MOS_j) and PLTs ($\bar{t}_j$) at web page j can be expressed as:

$$MOS_j = \frac{1}{n_j} \sum_{i=1}^{n_j} OS_{i,j} \qquad (1)$$

$$\bar{t}_j = \frac{1}{n_j} \sum_{i=1}^{n_j} t_{i,j} \qquad (2)$$

Where $OS_{i,j}$ represents opinion of user i at web page j and n_j represents total number of opinion scores received for web page j.

As shown in Figure 3, generally user QoE plunges as soon as PLTs ascend. For example, in the first session S1, PLTs are below 1 s while MOS is above 4.5. However, when PLT increases to 3 s at page 1 of session S2, MOS drops to 4.1, immediately. This indicates how subjects notice increasing delays of various intensities across all their shopping sessions. In contrast to sharp fall of QoE in the case of increasing PLT, the QoE does not increase as sharply in the case of decreasing PLT. Although, subjects usually express their contentment in the form of higher ratings as soon as waiting times descend, they still abstain from resorting to same ratings as those observed during the pre-disturbance period. This shows that the memory or recency effect prevails among the subjects, this phenomenon becomes quite evident by noticing the values of MOS for web pages 5–6, and 17–18. Obviously, the MOS values do not return back to the same level as appears before encountering additional delays on page 4 and page 16, respectively. Moreover, the MOS values are 4.6 and 3.6 on pages 1 and 36, respectively, showing a significant loss in QoE over time, despite of similar PLTs (less than 1 s). This is further illustrated in Figure 4. The ratio of subjects giving opinion score of 5 (Excellent) decreased significantly for page 36 (coming back from disturbances). At page 1, the average PLT was around 0.7 s and the share of rating 5 was more than 75%, while, at page 36 the average PLT was about 0.2 s and the share of rating 5 decreased to less than 20%. This significant drop in rating level shows the impact of accumulation of waiting time effects in the users' working memory, manifested in the recent past. Similarly, in paper [10], Hossfeld et al also showed the impact of memory effect on user QoE.

Additionally, we also observe that the QoE recovers significantly when no network disturbances occur during the whole task (shopping session). This is depicted by MOS for pages 13–15 (session S5) and pages 25–27 (session S9). The MOS of these sessions are approximately similar to the MOS of session S1. In contrast, if users face network disturbances on one of the pages during a shopping session (task), their ratings for subsequent pages of that respective session remain significantly low. This can be witnessed, for example, by observing the MOS levels of S2, S6 and S10.

In order to further strengthen our understanding of the observed decay in MOS and underlying memory effect among users, we computed standard deviation of opinions scores for
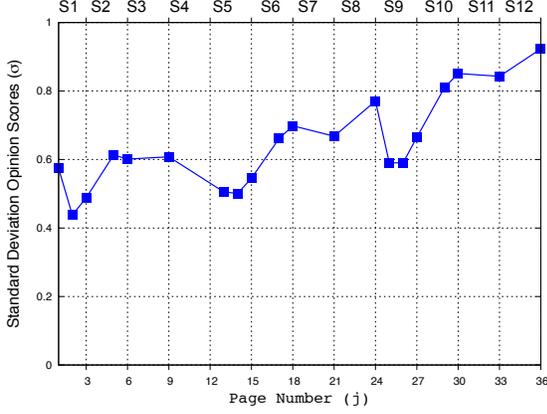
Fig. 5. Standard deviation in opinion scores of subjects for undisturbed web page transfers

web pages, where no network disturbances were applied. The PLTs for these web pages were kept well below 1 s. The objective was to determine whether the standard deviation varies significantly over time throughout the course of the experiment. Let σ_j represents standard deviation of opinion scores at page j, which is expressed by:

$$\sigma_j = \sqrt{\frac{1}{n_j} \sum_{i=1}^{n_j} (OS_{i,j} - MOS_j)^2} \qquad (3)$$

Where j represents only those web page transfers, which are not disturbed by network emulator, yielding page load times were below 1 s.

Figure 5 shows the standard deviation of opinion scores for web pages with undisturbed network settings (PLTs below 1 s). The difference of opinion scores among subjects gradually increase as illustrated by the increasing standard deviation. This indicates that the memory effect is subject to change among individuals. Some users tend to resist more than others in reinstating to their perception about service quality. It was probably the reason why there was increasing standard deviation.

B. Segmentation of users

As shown, increasing standard deviation expresses the existence of dispersion in opinions about the service quality. Therefore, it is imperative to segment the subjects into different categories before interpreting their responses. Before segmentation, we performed linear regression between PLTs and opinion scores for each subject. The reason for performing linear regression is to determine how each subject adopts her opinion score with change in PLT.

Let QoE_i and t_i denote the opinion scores and PLTs received from the user i, respectively. Similarly, let α_i and β_i be the intercept and slope of the equation for user i, respectively. Then after applying linear regression on opinion scores and PLTs of user i, we got the following equation:

$$QoE_i = \alpha_i + \beta_i \cdot t_i \qquad (4)$$

Hence, we extracted 44 pairs of α and β, each pair representing each of the 44 subjects. In order to divide subjects into segments, we performed clustering by applying the popular k-means clustering algorithm [17] on the values of α and β. Before applying k-means clustering on the set of α and β, optimal number of clusters k needs to be determined. We performed the following steps to determine the number of clusters:

1) Set k=2.
2) Apply k-means clustering with input k. Extract cluster centroids (μ) for each of the clusters.
3) Compute sum of squares of differences between data points in a cluster and their respective μ. Let D_j be the sum of squares of differences within cluster j. Let the centroid of cluster j be represented by μ_j. Moreover, let x_{ij} be a data point i in cluster j, and m_j represents the total number of data points in a cluster j. Then, D_j can be expressed by:

$$D_j = \sum_{i=1}^{m_j} (x_{ij} - \mu_j)^2 \qquad (5)$$

4) Add values of all D_j and compute the total sum of squares of differences (D) for all clusters k, as follows:

$$D = \sum_{j=1}^{k} D_j \qquad (6)$$

5) Plot values of D against corresponding values of k.
6) Repeat steps 2 to 4 after incrementing k by 1, for an arbitrary number of times n until elbow can be seen in the plot. We performed these steps by setting values of k to 2,3,4,5 and 6.

The plot of D versus k is shown in Figure 6. The elbow can be observed for $k = 4$. The reduction in D becomes marginal once k goes above 4. Hence, we set $k = 4$ in our study and applied k-means clustering. We obtained 5 users in cluster one, 13 users in cluster two, 19 users in cluster 3 and 6 users in cluster 4. The values of α and β for users within each cluster are depicted in Figure 7.

Figure 8 presents MOS per web page per cluster of users. From the plot, it becomes evident that the MOS of each cluster is following almost a similar pattern. As soon as PLTs increase, their respective MOS decrease steeply. However, when PLTs decrease, all clusters show a sign of memory effect and therefore, their respective MOS grows grudgingly.

Generally, subjects in cluster one appear to be the most tolerant as compared to subjects in any other cluster. Their MOS at any stage of the experiment does not go below the opinion score of 2. This infers that in comparison, this type of users is more optimistic about service quality and hence, can be a source of positive words of mouth for the service provider. On contrary, subjects in cluster four show the most negative response to delays. Their opinion scores go already below 3 as PLTs approach 3 s. After being exposed to network
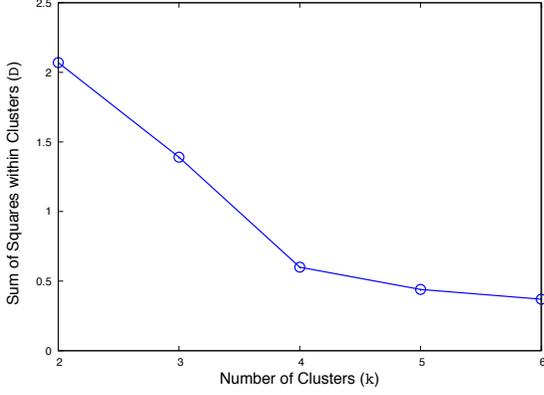
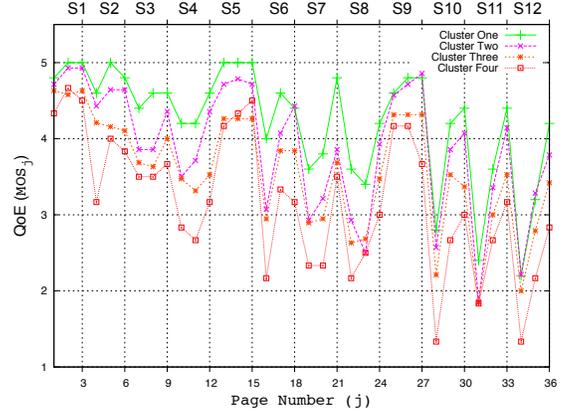Fig. 6. Elbow criterion: Sum of squares versus number of clusters
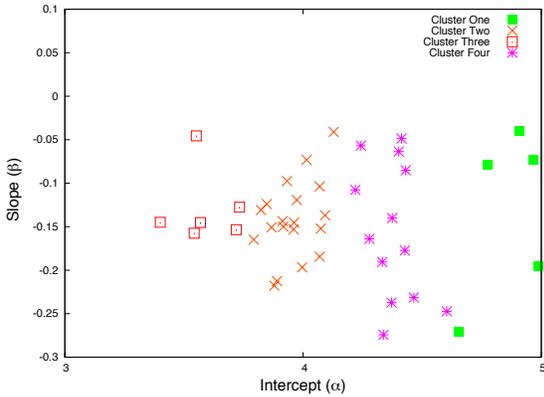

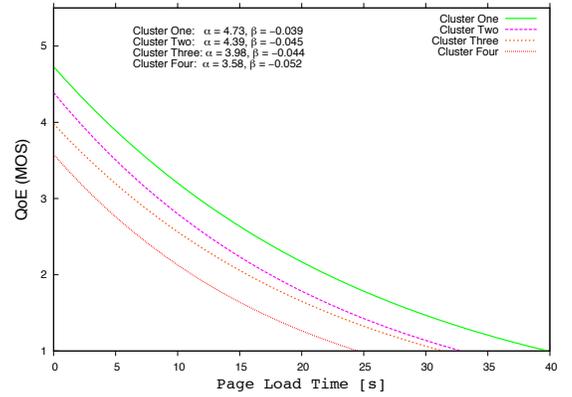Fig. 8. MOS of user clusters

Fig. 7. Clusters of users

Fig. 9. Exponential fits of QoE and PLTs of clusters

disturbances, these subjects hardly return to their initial level of satisfaction. Retention of such users can be challenging for service providers. Cluster two and three show a rather moderate and stable behavior. These two clusters accumulate to form the biggest segment of subjects that participated in our experiments. The stability in the opinion scores of these subjects show a strong indication of the memory effect. They tend to stay firm about their opinions despite of variations in the PLTs. Nevertheless, a sharp contrast is evident between cluster one and cluster four.

Self-herding behavior: As evident from Figure 8, the MOS of each cluster across the sessions S2 to S12 remains below their corresponding MOS of session S1. It shows that the users usually stick to the decisions taken by them in the recent past and therefore, do not rate the service quality above the level of rating provided in the first session. This behavior can be explained by a terminology called "self-herding" [18]. Self-herding refers to the tendency of a person to follow consciously or subconsciously her own decisions taken in the past.

Type A/B behavior pattern: In the study [19], the authors classify humans into two broad categories based on their

tolerance to delays: type A and type B. Type A personality users are rather impulsive, time urgent and aggressive as opposed to Type B that are patient, focused and easy-going. In our study, we observed some shades of Type A/B personality as a contrast of impulsion versus patience encompassing cluster one to four (i.e. cluster 1 $\cong$ rather A and cluster 2 $\cong$ rather B).

Finally, we tested multiple linear regressions on average PLTs and their respective MOS values for each of the clusters, separately. Exponential regression (cf. Equation 7) appeared to fit best on the data of each cluster with Pearson Correlation (r) values equal to -0.91, -0.90, -0.88 and -0.80 for cluster one, cluster two, cluster three and cluster four, respectively. Figure 9 presents plots of exponential best fit for each of the clusters, with their corresponding α and β values.

$$QoE = \alpha \cdot e^{\beta \cdot t} \qquad (7)$$

Vierordt's law: It is evident from Figure 9 that when PLTs are below 2 s, any increase in PLTs result in faster decay in MOS. However, when PLTs are higher than 6 s, the decay in MOS becomes rather slow. This observation can be

explained by Vierordt's law [20]. According to the latter, users either overestimate or underestimate the duration of the delay. Usually, they tend to overestimate durations of less than 2 seconds, accurately estimate durations between 2 to 6 seconds, and underestimate durations when delays are above 6 seconds. This law was further confirmed by an experimental study [21]. Similarly, the exponential interdependency of QoE and QoS was presented with the term IQX hypothesis by Fiedler et al. [9], which stated that, if the QoE is very high, a slight disturbance will decrease QoE strongly. On the contrary, if the QoE is already low, slight disturbance will not reduce QoE much. In this regard, one can relate the IQX hypothesis to Vierordt's law using the results presented in this study. The small PLTs (less than 2 s) imply high QoE, where any slight increase in PLTs are overestimated by users, bringing QoE down rather strongly.

V. CONCLUSION

In this paper, we presented a set of observations with regards to the impact of short network outages (OFF times) on QoE. The results were obtained by performing a lab study, which involved user subjective tests in the context of web browsing. There are several conclusions extracted from this study, which are discussed below.

First, the impact of recurring OFF times on QoE accumulate with time, which can be seen in the form of decaying trends in the user opinion scores. When the waiting times grow, user satisfaction level drops immediately. Subsequently, user opinions do not witness the sign of complete recovery after the network problems are rectified. This is an evidence of the role of memory effect and time, which need to be considered as important variables in the modeling of QoE. Second, we found the existence of four different user segments with regards to their feedback against varying page load times. While all these segments witness memory effects, some users are rather impulsive than others in changing their opinions. Hence, their opinion scores drop rather steeply with increase in page load times. Nevertheless, we witness an exponential relationship between QoE and page load times in all cases, regardless of the user segment. Third, on average, the opinion scores of each segment of users remain under the level of MOS given during their corresponding first shopping sessions before the occurrence of disturbances. Specifically, their scores do not seem to rise above their respective initial MOS levels.

This is ongoing work, which includes the assessment of the reaction of users to varying frequencies and durations of network outages. We are trying to model how users experience a typical (mobile) network service over time and the associated factors that can influence their feedback about that service. Moreover, we intend to assess the impact of disturbances within different key phases of a transfer from the network perspective. Another important aspect of our work is the extension of these user tests beyond the lab to the field settings.

REFERENCES

[1] T. Qiu, J. Feng, Z. Ge, J. Wang, J. Xu, J. Yates. "Listen to me if you can: Tracking User Experience of Mobile Network on Social Media," *In Proceedings of the 10th ACM SIGCOMM Conference on Internet Measurement.* pp.288–293, Melbourne, Australia, 2010.

[2] J. Shaikh, M. Fiedler, P. Arlos, D. Collange. "Modeling and Analysis of Web Usage and Experience Based on Link-level Measurements," *In Proceedings of the 24th International Teletraffic Congress.* pp. 1–8, Krakow, Poland, September 2012.

[3] S. Egger, T. Hossfeld, R. Schatz, M. Fiedler. "Waiting Times in Quality of Experience for Web Based Services," *In Proceedings of the 4th International Workshop on Quality of Multimedia Experience.* pp. 86–96, Yarra Valley, Australia, July 2012.

[4] N. Bhatti, A. Bouch, A. Kuchinsky. "Integrating User-perceived Quality into Web Server Design," *Computer Networks.* Volume 33, Issue 1–6, pp. 1–16, 2003.

[5] C. Lorentzen, M. Fiedler, H. Johnson, J. Shaikh, I. Jorstad. "On User Perception of Web Login–A Study on QoE in the Context of Security," *In Proceedings of Australasian Telecommunication Networks and Applications Conference.* pp. 84–89, Auckland, New Zaeland, November 2010.

[6] T. Hossfeld, S Egger, R. Schatz, M. Fiedler. "Initial Delay vs. Interruptions: Between the Devil and the Deep Blue sea," *In Proceedings of Fourth International Workshop on Quality of Multimedia Experience.* pp. 86–96, Yarra Valley, Australia, July 2012.

[7] J. Shaikh, M. Fiedler, D. Collange. "Quality of Experience from User and Network Perspectives," *Annals of Telecommunications.* Volume 65, Issue 1–2, pp. 47–57, 2010.

[8] T. Minhas, M. Fiedler. "Quality of Experience Hourglass Model," *In Proceedings of International Conference on Computing Management and Telecommunications.* pp. 87–92, Ho Chi Minh City, Vietnam, January 2013.

[9] M. Fiedler, T. Hossfeld, P. Tran-Gia. "A Generic Quantitative Relationship between Quality of Experience and Quality of Service," *IEEE Network.* Volume 24, Issue 2, pp. 36–41, 2010.

[10] T. Hossfeld, S. Biedermann, R. Schatz, A. Platzer, S. Egger, M. Fiedler. "The Memory Effect and its Implications on Web QoE Modeling," *In Proceedings of the 23rd International Teletraffic Congress.* pp. 103–110, San Francisco, USA, September 2011.

[11] J. Garcia, E. Conchon, T. Perennou, A. Brunstrom. "KauNet: Improving Reproducibility for Wireless and Mobile Research," *In Proceedings of the 1st International Workshop on System Evaluation for Mobile Platforms.* pp. 21–26, San Juan, Peurto Rico, 2007.

[12] "Usage Statistics and Market Share of Apache for websites," URL: *http://w3techs.com/technologies/details/ws-apache/all/all*, Last visited: June 2013.

[13] "CodeIgniter: A fully Baked PHP Framework," URL: *http://ellislab.com/codeigniter*, Last visited: June 2013.

[14] P. Arlos, M. Fiedler, A. Nilsson. "A Distributed Passive Measurement Infrastructure," *Lecture Notes in Computer Science.* Volume 3431, pp. 215–227, 2005.

[15] "Browser statistics," URL: *http://www.w3schools.com/browsers/browsers_stats.asp*, Last visited: June 2013.

[16] "Fiddler: The Free Web Dubugging Proxy for any Browser, System or Platform," URL: *http://fiddler2.com/*, Last visited: June 2013.

[17] J. Hartigan, M. Wong. 'Algorithm AS 136: A K-Means Clustering Algorithm," *Journal of the Royal Statistical Society.* Volume 28, Issue 1, pp. 100–108, 1979.

[18] D. Ariely, M. Norton. "How Action Create – Not Just Reveal – Preferences" *Trends in Cognitive Sciences.* Volume 12, Issue 1, pp. 13–16, 2008.

[19] M. Marquis, L. Dubt, J. Chebat. "Consumers' Response to Waiting Time: New Segmentation Bases are Required for Service Industries," *Advances in Consumer Research.* Volume 21, pp. 548–553, 1994.

[20] K. Vierordt. "Der Zeitsinn Nach Versuchen," *Book published by H. Laupp.* 1868.

[21] E. Wang. "User's Delay Perception and Tolerance in Human-Computer Interaction." *In Proceedings of the Human Factors and Ergonomics Society Annual Meeting.* Volume 46, Issue 5, pp. 651–655, 2002.