In Small Chunks or All at Once? User Preferences of Network Delays in Web Browsing Sessions

Nazrul Islam, Vijaya John David Elepe, Junaid Shaikh, Markus Fiedler

First International Workshop on QoE-centric Network and Application Management (QoENAM)

2014 Sydney

In Small Chunks or All at Once? User Preferences of Network Delays in Web Browsing Sessions

Nazrul Islam, Vijaya John David Elepe, Junaid Shaikh, Markus Fiedler
Department of Communication Systems
Blekinge Institute of Technology
Karlskrona, Sweden
Email: {nais11, viel12}@student.bth.se, {junaid.junaid, markus.fiedler}@bth.se

***Abstract*—The time-critical tasks on the Internet are increasing. The delays in these tasks can have severe implications on the Quality of Experience (QoE) of a service. Therefore, networks require smart user-centric resource management mechanisms to reduce the impact of these delays on QoE. For this, a better understanding of the user preferences with regards to service performance is a prerequisite. In this paper, we present user responses to the three different distributions of delays, occurring during shopping sessions on the Web. By keeping the overall waiting time of the sessions same, we show how the users respond differently to the different set of delays. We analyzed the user responses and found that, the users prefer small frequently occurring delays as compared to the long rarely occurring delays within a task-based session.**

I. Introduction

People rely heavily on the wide domain of applications and services running on the Internet. Large number of these applications are mainly accessed via World Wide Web (WWW). Generally, the users expect a faster delivery of response for any request they make, without any disturbance. However, in the context of web browsing, disturbance in the form of end-user waiting time is a common occurrence and also the key determinant factor of Quality of Experience (QoE) [1]. The waiting time is defined as the time between a client sending a request to a server and the response to that particular request is fully visible to the client.

Despite of the dramatic increase in network bandwidth over the years, networks are still not smart enough to serve these web objects immediately, according to the user expectations. Particularly, the cellular networks may even take multiple of seconds to fetch small web objects from the web servers. Several reasons describe these delays.

First, the cellular channel quality varies significantly over time. The link rates change dramatically, which makes downloads bursty and thus, produces many short outages during the transmission of packets on the network.

Second, resources are shared among multiple users on the network. The scheduler managing these resources may sometimes take significantly long time to assign resources to certain transfers. As a consequence, packets may suffer from long waiting times in the queue.

Third, multiple transfers launched by a user simultaneously may suffer from short-term outages due to the self-inflicted delays. The traffic from multiple transfers may compete with each other. Consider a scenario when a user downloads a long file and at the same time performs web browsing. The short transfers of web browsing application may suffer long delays, as packets get stuck in potentially long queues at the gateways, due to the heavy traffic generated by the file download.

For the above reasons, QoE-based management of networks is gaining central importance in the success of services provided by a network operator. Network operators need to deploy efficient and user-centric resource management mechanisms. Obviously, they need to share the resources and due to the resource sharing, delays may occur. However, they need to be aware of the user preferences and thus, share these resources in a way that minimizes the impact of delays on the user QoE.

On this background, we evaluate the impact of duration and frequency of disturbances on the web browsing QoE of online shoppers. These disturbances appear in the form of packet delays. We create situations analogous to the cellular networks, and put certain packets in a queue at an intermediate node on the network. Either long delays appear all at once and then, problems get resolved, or short delays appear continuously for a long period of time. The long delays for a short period of time means, users suffer from a considerably high Page Load Time ($\sim$16 seconds) on a single web page in a session, and then all the other pages load normally without any additional delay (within less than 1 second). The short delays for a long period of time indicate a situation when a multiple number of pages in a session continuously suffer from the short additional delays ($\sim$4 seconds).

In this paper, we will evaluate how users respond to the above situations. The results of this study will propose a set of principles for the QoE-based performance management, which is an integral part of the functional dimension of network management, i.e., FCAPS (Fault, Configuration, Accounting, Performance and Security Management) [2].

Previously, a several number of papers reported the impact of delays on the end-user QoE [3], [4]. Another study showed that the user-QoE drops significantly over time when the Page Load Time grows but it does not recover completely after the network problem is resolved [5]. In [6], authors show that the user satisfaction level breaks when waiting times exceed 10 seconds in a single session. Similarly, another study illustrated the importance of short waiting times in the case of e-commerce services [7]. The users cancel download of images when the waiting time exceeds 10 to 20 seconds [1]. The study [8] that estimates for user tolerance of Quality of Service (QoS) for an e-commerce website states that, the

delay between 2–6 seconds can be estimated accurately by the users. All the related studies draw some thresholds on tolerable waiting times, based on the user QoE. According to the best of our knowledge, there is no study, which describes the trade-off between the duration and frequency of delays in a systematic manner. The next section describes our methodology further in detail.

The remainder of this paper is structured as follows. Section II provides details about the research methodology. Section III presents experiment setup used in this study to conduct user tests. Section IV analysis and results from the experiment. Finally, Section V poses a set of summaries and concludes with future work.

II. Methodology

In this study a total of 42 participants took part in the experiment. The mean age of all participants was 26 years. The maximum age was 33 and a minimum age was 19 years. All these subjects were regular users of Internet and use e-commerce website for online shopping. We provided a primary training session of 5 minutes for each subject. During the training all necessary instructions required to perform the test are provided. A task-driven process was provided. These tasks were selecting a category, then a product in the category and purchasing a product in it. We provided three shopping sessions for each subject. Each session was defined with a browsing of five web pages: Category selection page, product selection page, product details page, payment details and payment confirmation page.

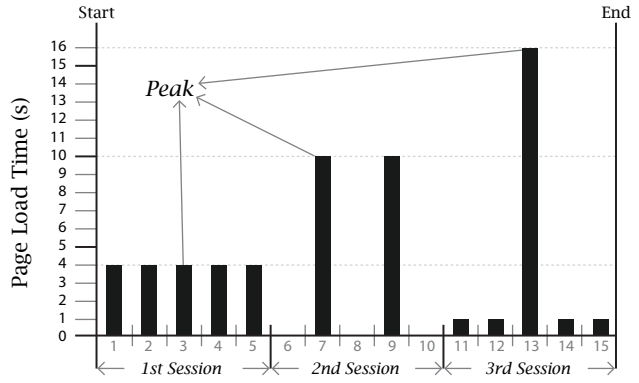

Fig. 1. Web session with different *peak* delays

Figure 1 represents one complete experiment of a subject, where 1st, 2nd and 3rd session represents three individual shopping sessions with the *peak* delays of 4 seconds, 10 seconds and 16 seconds, respectively at any web page of the session. In this paper, the term *peak* delay represents the highest Page Load Time faced with a subject during a shopping session of five web pages. In Figure 1, the y-axis represents the Page Load Time for each individual web page, which indicates the amount of delay perceived by the subject for each web page. Despite the difference in the delay pattern, we kept the total waiting time of every shopping session approximately 20 seconds. We want to see whether the users report their experience differently due to the different *peak* delays within a session?

The applied delay patterns can be viewed from Figure 1. The first session represents 4 s session, in which we put a continuous delay where users perceived continuously 4 seconds of delay on all pages in the session. Second session represents 10 s session, in which we put a delay in the second and the fourth pages in which user perceived a delay of 10 seconds on both these pages. The third session represents the 16 s session, where we put a large delay on the third page, where all users perceived a delay of 16 seconds on this page in the web session.

Additionally, we randomized the order in which the above mentioned sessions appear to each subject. Some users had 4 seconds of *peak* delay in the first session, followed by 10 seconds of delay in the second and then the 16 seconds delay at the end. While some other users experienced 16 seconds of delay at the beginning, followed by 4 seconds of delay and then 10 seconds of delay at the end. All the delay patterns are mentioned in Table I. Hence, every subject experienced these sessions with a random occurrence.

At the end of each purchase session, subjects were asked to answer these following questions:

1) *How do you feel about the overall loading time?*

   The options for this question were provided based on the five points ACR scale for rating quality, which is recommended by ITU-T [9].

2) *Would you be willing to use this internet service again?*

   The options for answering this question were given as follows:

   - Yes
   - No

Based on the results obtained from these questions from each of the subjects, a detailed analysis is made to find out the impact of these different *peak* delays.

III. Experiment Setup

In order to find out the effect of the network disturbances on a web browsing session, an experimental setup was established having a server, a client and a network emulator. The network emulator was placed between the server and the client to generate desired network environment (Figure 2). The KauNet [10] was installed and configured in Linux environment (Ubuntu 10.04) as the network emulator. All traffic passes through the network emulator and the bandwidth of 10 Mbps link.

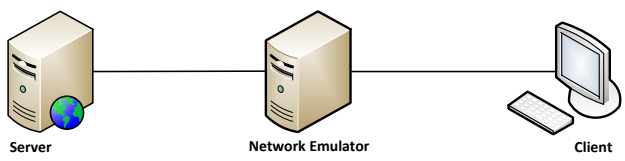

Fig. 2. Experiment setup

We used popular web server called Apache web server (Apache 2.2) [11] configured on the server machine. The application Bind9 [12] is installed and configured on the server for Domain Name System (DNS) service. Server machine was

setup with Ubuntu 10.10 and client machine with Windows 7 operating systems. All the web pages on web server were deployed using a well known PHP framework called CodeIgniter [13]. CodeIgniter is an open-source, lightweight, powerful web application framework. These web pages were accessed by the client side web browser (Google Chrome) based on the request. Where Google Chrome browser was set to Incognito mode [14]. The abduction of all HTTP(s) traffic between the client computer and the Internet over the Windows platform was done by an open-source web debugging proxy tool known as Fiddler [15].

We used fiddler to collect the logs on the client side. These logs were stored in HTTP ARchive (HAR) files. A PHP script was used to extract all the required information presented in the 'har' file. This script fetched timestamps based on the first request from the client and the last response from the server. To collect the network level trace, we used network protocol analyzer T-shark [16], which was used on the client machine. It captured all the packets from the client-server communication on the network-level. All files are stored locally in a 'pcap' format. A Perl script was developed to extract the timestamp of the request from the client and the last response being sent by the server to the client.

We developed an automated tool to manage the entire experiment and placed on the client machine. The end-user can have a continuous flow of a real life web browsing experience without interruption. Based on the design of the experiment process (User ID, Session ID and URL), this script fetches the desired network settings specified by the user for that session and signaled the network settings to the network emulator. Furthermore, it collects answers to the question, mentioned in the earlier section. This information was stored in the local database (MySQL) [17] based on the User ID, Session ID, network settings and answers to the questions.

IV. Analysis and Results

A. Impact of peak delays on end-user QoE

This section discusses the overall impact of *peak* delays in a session. As we mentioned in the previous section, *peak* delay refers to the highest Page Load Time faced by a user within a shopping session. The users performed three shopping sessions and thus, went through three different peak delays, which we induced systematically in a controlled manner. However, the total waiting time, i.e., the accumulated Page Load Times of five pages for each of the three sessions was the same, i.e., 20 seconds.

Figure 3 illustrates the Mean Opinion Score (MOS) values given by the users in each of the three sessions. Clearly, the users prefer delays to be short, even if these delays continue to occur for many consecutive web pages during a session. They gave better ratings in the session in which all the five web pages loaded in 4 seconds (approx. MOS value of 3.2), in comparison to the session in which one web page took 16 seconds to load, while, all the other four web pages loaded within 1 second of time (approx. MOS value of 2.5).

Additionally, we also report the share of user ratings for each of the sessions. Figure 4 displays bar chart of ratings for each session. The session with 4 seconds of peak delays leads

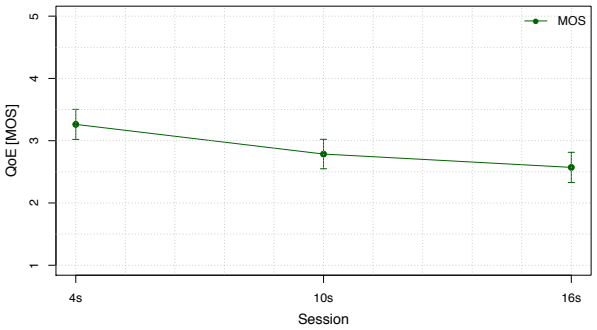

Fig. 3. Average user satisfaction for three different delay session

to more positive percentage ratings, as compared to the other two sessions. Note that, none of the users reports "Excellent" rating. Obviously, the peak delay was not less than 4 seconds in any of the sessions. Therefore, the users noticed delay and their flow of thoughts got interrupted during the task, which motivated them to provide ratings below "Excellent". The "Good" and the "Fair" ratings constitute of more than 80% of the total ratings for 4 seconds of *peak* delay. This share reduces significantly for the sessions with peak delays of 10 seconds and 16 seconds, and is replaced mostly by the climbing "Poor" ratings.

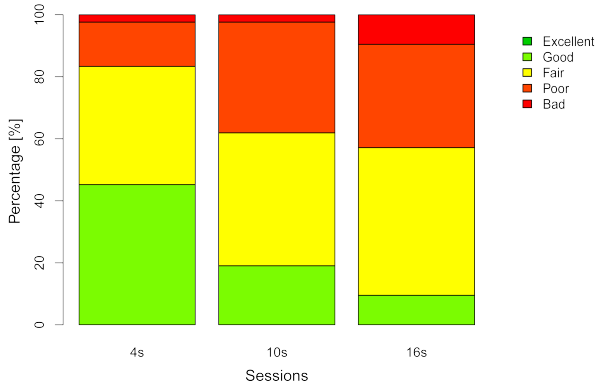

Fig. 4. User satisfaction for three different delay sessions

The above findings suggest that users do not like long network disturbances, occurring all at once. The users are more tolerant towards low intensity network disturbances continuing over a long period of time. Although, the accumulated waiting time in every session is approximately the same, but the amount of *peak* delay makes the difference in the user given MOS. Obviously, the users avoid situations where they have to wait for too long without any response. The network resources management mechanisms should ensure that, the end users keep getting at least some piece of response, without waiting for too long at a time. The end-user waiting times in one chunk should therefore be small and distributed over time.

B. Impact of peak Delay on End-User Acceptability

In this subsection, we present the impact of *peak* delay on the end-user acceptability of service. We provided users the option to express their acceptability of the delays by answering "Yes" or "No" to a close-ended question, which is mentioned in the previous section. It is interesting to see a suggestive difference between their responses for different sessions, which are displayed in Figure 5.

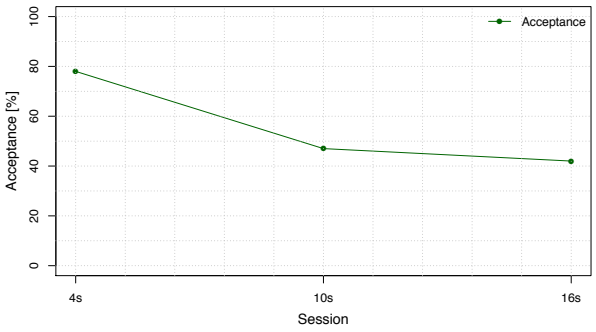

Fig. 5. User acceptability ratings for different *peak* delay

Around 80% of the users find a continuous 4 seconds of Page Load Time acceptable to them. However, more than half of the users find the service unacceptable, if any one web page during a shopping session takes more than 10 seconds of delay to load. Once again note that, the overall waiting time of the whole session is 20 seconds, for each of the three sessions, but the end-user acceptability is not the same due to the difference in peak delays (the highest Page Load Time). Network operators need to make sure that the network service does not disappear continuously, for a long period of time, particularly, when a user is actively using an interactive application like web browsing.

C. Impact of delay sequence on QoE

This section discusses the impact of the order in which sessions appear. As mentioned previously, we randomized the sequence in which users faced delays. In the experiment, a user could go through one of the six possible sequences of sessions, which are mentioned in the Table I. For example, shopping session with the *peak* delay of 4 seconds may appear at any of the three possible positions: the first session (start), the second session (mid) or the third session (end). This randomization of the order makes sure that our results do not get biased by the sequence in which delays occur.

TABLE I. COMBINATIONS OF SEQUENCES FOR DIFFERENT PEAK DELAY SESSIONS

Sequence of order	Sesssion Appearance
1	4 s, 10 s, 16 s
2	4 s, 16 s, 10 s
3	10 s, 4 s, 16 s
4	10 s, 16 s, 4 s
5	16 s, 4 s, 10 s
6	16 s, 10 s, 4 s

In Figure 6, the x-axis represents a session's position of appearance to a user during the experiment and y-axis represents the user given MOS. The "start", the "mid" and the "end" indicate the location of each session. We observe that the MOS for 4 seconds peak delay session remain significantly higher than the other sessions regardless of the position. We do not find a clear trend due to the location, but it is clear that, the users prefer small disturbances appearing frequently, as compared to a long disturbance appearing without any response.

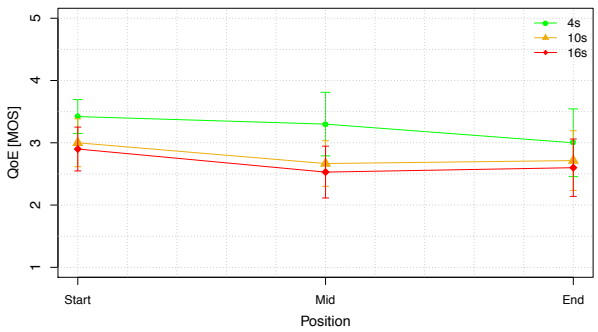

Fig. 6. MOS at different positions for all three sessions

From these observations, we can point out that, though the overall waiting times are approximately same for each session, the user given MOS is different depending on the position where delay has been perceived. The later scores for a same amount of delay do not seem to rise above the respective initial MOS score.

D. Network-level analysis

In this subsection, we provide a brief look at events, occurring at the packet-level. It is interesting to look at the reactions of the protocols, when we induce certain delays on the packets. We targeted packets transmitted from the server to the client, carrying response of the object requested by the client. Thus, we created conditions in which the packets get stuck in a queue at the intermediate node.

Generally, on average, two TCP connections open at the first page of a session. These connections do not usually terminate at the end of page download. They continue over the next web pages of the session and entertain further subsequent requests from the client. Therefore, we do not observe terminations of TCP connections, until the end of the shopping session in normal scenario. However, when hold packets in a queue, we start to see the abnormalities, which are described below.

When we apply delay on the packet carrying the response from the server, we notice that the client keeps initiating new TCP connections with "SYN" packets, frequently, until it receives packets from the server. The higher the amount of delay, the greater the number of TCP connection requests from the client. For example, when we applied approximately 10 seconds of delay on a packet from the server side, we found 9 additional SYN requests from the client side, as mentioned

in Table II. Similarly, we observe a large number of TCP terminations initiated by the server, when we apply delays higher than 10 seconds. The exact number of TCP connections termination are listed in Table III. When the server does not get ACKs of the sent packets, it starts terminating the existing TCP connections, and keep on retransmitting the FIN packets.

TABLE II. NUMBER OF TCP CONNECTION INITIATIONS AT ONE OF THE WEB PAGES. HIGHER NUMBER OF TCP CONNECTION INITIATIONS FOR HIGHER DELAYS

Delay	Mean	STD	95% CI
No delay	3.41 s	1.17 s	10.40%
4 s	5.43 s	1.85 s	10.31%
10 s	12.33 s	1.43 s	3.50%

TABLE III. TCP CONNECTION TERMINATIONS OBSERVED AT ONE OF THE WEB PAGES. LARGE NUMBER OF TERMINATIONS FOR 16 S DELAY

Delay	Mean	STD	95% CI
No delay	0	0	
4 s	0.15 s	0.48 s	100%
16 s	12.03 s	4.24 s	10.65%

Similarly, the client also retransmits the object requests by sending GET requests for the same object recursively. We observe a significant increase in the amount of GET requests from the client side when we apply delay on packets.

The opening of many sockets for a single web page download wastes lot of resources at both end systems, and also produce unwanted traffic on the network. Such delay events do not only waste resources but also produce a multiplicative impact on the end-to-end performance of data transfers as well as on the QoE.

V. QOE-BASED NETWORK MANAGEMENT

A user session on the network is a continuous flow of experience, where a user comes across a series of waiting times. These waiting times are caused by the end-to-end packet delays. At times, the packet delays are so high that they damage badly the QoE of whole session. Over the years, it has been witnessed through many studies that the major reason behind such delays is the inappropriate management of network resources, and more specifically, the inefficient adaptation of link capacity. The assignment of link capacity need to be user-centric and takes into account the strategies, which maximize the overall QoE of a user session.

The outcomes of this study show that the users prefer small waiting times over sudden long delays within a session. The results provide a guideline to network operators for the user-centric management of link capacity. The link capacity adaptation mechanisms must ensure that a user at least gets a minimal level of service without long interruptions, such that, a constant flow of requested data is received by the user. Specifically, network operators should avoid the assignment of link capacity, which is much higher than the amount of anticipated consumption, followed by multiple seconds of outages. Instead, they need to shape link capacity based on user QoE, such that, it ensures constant flow of response. The user-centric network management will thus help network operators to improve customer satisfaction and reduce underutilization of the available link capacity.

We observed in this study that the users do tolerate waiting times in a session, if they are distributed in small chunks over times. Our results show that a moderate link capacity resulting in constantly occurring waiting times around 4 s are more acceptable to the users in comparison to the high link capacity (waiting times below 1 s), followed by occasional long waiting times. Moreover, our results also showed how the inefficient assignment of resources result in the abrupt initiation and terminations of TCP connections, which further result in the degradation of performance as well wastage of network and system-level resources. Hence, the results in this study signifies the importance of user-centric adaptation of link capacity and more generally, user-centric, QoE-based management of networks.

VI. CONCLUSION

In this paper, we presented the results of our study on the user responses to delay distributions during task-based shopping sessions. Our study was based on the subjective experiments performed in lab environment with users. In this study, we designed three different test conditions for each user. Each condition refers to a specific type of peak delay induced during page loading process. The main question of this study was to determine whether users like long delays occurring all at once or they prefer short delays, occurring continuously on many consecutive pages. In the experiment, each user went through three web browsing sessions to perform shopping tasks. Each session was based on 5 web pages. In one session, they faced 4 s of additional delay on each of the five pages. In the another one, they went through 10 s of additional delay on two pages, while no delay at the remaining three pages. In the third session, they faced a long delay of 16 s all at once on one of the web pages, while the other four web pages had no additional delay. We found out that the users prefer short delays (4 s) occurring continuously, in comparison to the rarely occurring long delays. Their MOS score was above 3 for 4 s of peak delay, which reduced to 2.5 for 16 s of peak delay. Users also found the service unacceptable, when they faced 16 s of delay on one of the web pages during a session. Although, the cumulative waiting time of each session was 20 s, the session with low peak delay obtained better ratings from the users.

Network operators may use these results to devise better user-centric strategies for resource management. In particular, they need to take into account that the interactive activities like web browsing, and particularly the shopping tasks on Web, require immediate responses from the server. Any additional delays due to channel scheduling, long packet queues or link quality changes may produce significant impact on the user QoE. Therefore, the appropriate resources should be assigned to this type of application in order to keep the delays short. One of the considerations is to schedule the traffic such that the short time slots be assigned frequently in order to avoid long interruptions during the flow of data.

VII. FUTURE WORK

In order to understand the impact of variety of delay distributions on packets during a usage session, this research work needs to be extended to the real-life scenario using field trials or crowdsourcing experiments. The real-life scenario will allow more freedom to test the impact of variety of network

conditions on QoE. It is crucial to identify the impact of time-varying network performance on the overall session QoE of a user. Specifically, the studies need to quantify and model the impact of duration and frequency of extra network delays on user session QoE. This can be achieved by creating controlled time-varying network conditions using model-based network emulations. The future work also needs to extend this study to other popular applications and content types. The aim is to come up with a generalized model, which enable network operators to carry out QoE-based network management tasks, and thus, improve the customer satisfaction.

ACKNOWLEDGMENT

This research is supported by Vinnova through the Celtic-plus project Quality of Experience Estimators in Networks (QuEEN).

REFERENCES

[1] S. Egger, T. Hossfeld, R. Schatz, and M. Fiedler, "Waiting Times in Quality of Experience for Web Based Services," *in Fourth International Workshop on Quality of Multimedia Experience (QoMEX)*, pp. 86–96, Yarra Valley, Australia, July 2012.

[2] "ISO/IEC 10040:1998 - Information technology – Open Systems Interconnection – Systems management overview," http://www.iso.org/iso/iso_catalogue/catalogue_tc/catalogue_detail.htm?csnumber=24406, [Online; Accessed: 6-Jan-2014].

[3] J. Shaikh, M. Fiedler, and D. Collange, "Quality of Experience from User and Network Perspectives," *Annals of Telecommunications*, vol. 65, no. 1-2, pp. 47–57, 2010.

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

[5] J. Shaikh, M. Fiedler, P. Paul, S. Egger, and F. Guyard, "Back to Normal? Impact of Temporally Increasing Network Disturbances on QoE," *IEEE Workshop on Quality of Experience for Multimedia Communications*, Atlanta, USA, December 2013.

[6] I. Ceaparu, J. Lazar, K. Bessiere, J. Robinson, and B. Shneiderman, "Determining Causes and Severity of End-User Frustration," *International Journal of Human-Computer Interaction*, vol. 17, no. 3, pp. 333–356, 2004.

[7] A. Sieminski, "Changeability of Web Objects-Browser Perspective," *in Fifth International Conference on Intelligent Systems Design and Applications*, pp. 476–481, Wroclaw, Poland, September 2005.

[8] N. Bhatti, A. Bouch, and A. Kuchinsky, "Integrating User-Perceived Quality into Web Server Design," *Computer Networks*, vol. 33, no. 1, pp. 1–16, 2000.

[9] "Methods for Subjective Determination of Transmission Quality," http://www.itu.int/rec/T-REC-P.800-199608-I/en, 1996, [Online; Accessed: 18-Jan-2013].

[10] "KauNet," http://www.kau.se/en/kaunet, [Online; Accessed: 26-June-2013].

[11] "Welcome! The Apache HTTP Server Project," http://httpd.apache.org/, 2013, [Online; Accessed: 21-Jul-2013].

[12] "Bind9- Debian WikiBind9," https://wiki.debian.org/Bind9, 2013, [Online; Accessed: 18-Jul-2013].

[13] "CodeIgniter / EllisLab," http://ellislab.com/codeigniter, [Online; Accessed: 28-July-2013].

[14] "Chrome Browser: Incognito mode (browse in private)," https://support.google.com/chrome/answer/95464?hl=en, 2013, [Online; Accessed: 10-Jul-2013].

[15] "Fiddler: Web Debugging Proxy," http://www.fiddler2.com/fiddler2/, [Online; Accessed: 26-Aug-2013].

[16] "T-shark: The Wireshark Network Analyzer 1.10.0." http://www.wireshark.org/docs/man-pages/tshark.html, 2013, [Online; Accessed: 16-Nov-2014].

[17] "MySQL: The Worlds Most Popular Open Source Database," http://www.mysql.com/, 2013, [Online; Accessed: 14-Aug-2013].