Relationship between Quality-of-Service and Quality-of-Experience for Public Internet Service

Stas Khirman (stask@narus.com) and Peter Henriksen (peterh@narus.com)

Narus Inc. (http://www.narus.com)
3950 Fabian Way, Palo Alto, CA 94303

***Abstract*--To successfully resolve the Internet infrastructure's most pressing problems the Internet industry needs to improve the service quality. However, measurements and provisioning of the quality of service (QoS) are generally defined in terms of network delivery capacity and resource availability, not in terms of satisfaction to the end-user. The fundamental assumption behind such traditional provisioning is that the measured quality of service is closely related to the quality of experience (QoE) for the end-user. Through the use of Narus implementation of the Semantic Traffic Analysis technology we have been able to directly measure quality of experience for the user and establish the relationships between QoS and QoE.**

A. INTRODUCTION

The Internet industry is spending significant amounts of money to provide end-users with more networking resources so as to improve their satisfaction with the service. However, the relationship between deployed resources and end-user satisfaction is far from obvious. It is not clear how much is gained by improvements in delivery bandwidth or lower latency.

Our investigation is trying to relate the objective network service conditions with the human perception of the quality of the service. This subject has been widely investigated for voice delivery [1,2] and it is widely acknowledged that the relationship between voice transmission conditions and the human perception of quality is far from linear.

We discuss in detail how the human satisfaction of HTTP service (web browsing) is affected by the two main network QoS parameters, namely network delivery speed (bandwidth) and latency. Our goal is to measure the level of user dissatisfaction with the web-content delivery quality, find the most important component of this dissatisfaction and give recommendations regarding possible ways of improving the web service.

B. DATA COLLECTION

The most popular delivery mechanism for the Internet, the HTTP protocol, is implemented using a simple text based request/response approach where a client submits a request for a specific object (specified via the Uniform Resource Locator - URL) and the server then returns the appropriate response. The original HTTP/1.0 [3] protocol requires separate TCP connections to be established for every request/response pair, where end-of-delivery of the response object is signaled by server-side closing of the transport connection. The HTTP/1.1 [4] protocol permits TCP connections to be reused, implying that the server has to inform the client about the size of the response object to be transmitted through the "Content-Length" attribute in the HTTP server response header. Under normal circumstances, the underlying TCP connection will be gracefully closed, either by server or client, after the last request has been succesfully serviced.

Unsatisfactory conditions of the web delivery service could make the end-user push the STOP or RELOAD buttons in his web browser, initiating user-side cancellation. User-side request cancellation would result in early closing of the transport connection. This cancellation request can be identified by analysis of the HTTP protocol interactions. While user-side cancellation activity is not necessarily the only way to gauge a user's dissatisfaction with web content and/or underlying transport quality, it may be considered the only way to get an independent and unbiased measure.

Recent web analysis publications use a number of approaches to collect HTTP interaction information. This includes gathering information from HTTP server log files [5], HTTP cache server log files [6] and special instrumentation of the web browser [7].

Our intent was to investigate the behavior of unbiased users accessing public Internet services; unbiased in the sense that their behavior should not have been altered by the fact that information was being collected and analyzed. Crucial to the quality of such an analysis is the position of the data collector. We had no means of installing collection agents on the client side, nor did we have any means of predicting what web servers the user would visit. We considered using HTTP proxy log files/firewall information, but this would have contradicted our intention of collecting data for general public Internet access since firewalls are usually installed at companies only. Furthermore, end-users belonging to the same company often have a common interest that may introduce a bias into the results.

We instead used the Semantic Traffic Analysis (STA) methodology developed by Narus Inc.for data collection. STA is essentially a natural extension to passive network analysis (e.g. network sniffing, see RMON [8] and NetFlow [9]). The basic Narus STA component, the STA Analyzer, is a passive network-attached sniffing device that collects packets traveling across a specific network link. The STA Analyzer is applying “reverse engineering” to the captured packets. This involves keeping state of each individual TCP connection, reordering packets so they occur in correct sequence and applying protocol specific parsing so as to extract details of the application layer transactions.

The STA Analyzer has access to all non-scrambled information available to the application client and server. Specifically for HTTP protocol, STA is capable of restoring the full sequence of request/response transactions and report information which includes (but is not restricted to):

- client and server IP addresses
- requested URL
- response status
- response content type (As defined by “Content-Type” attribute)
- response advertised size (if any) (As defined by “Content-Length” attribute)
- actual size of the delivered object (or in case of cancellation the number of bytes transmitted)
- time elapsed from first request packet to first response packet
- time elapsed from first response packet to last response packet

It is important to note that the last three parameters are crucial for our analysis and can only be collected using network based mechanisms.

We installed our STA collector in a commercial ISP network that provides general public Internet access, so as to capture representative statistics for public Internet usage. We assume that the collected information is a good representation of unbiased customer behavior. A majority of the traffic (80 %) was generated by customers using dial-up modem connections (up to 56kBit/s). The rest of the traffic was generated by corporate users with high-speed connections and external requests to the hosted web servers. No HTTP cache servers were, to our best knowledge, deployed in the ISP network, so we feel it safe to assume that the collected HTTP information represents data retrieval from original remote web servers or remote distributed cache servers (such as Akamai). In both cases, we measure end-user experience of access to information over a wide area network.

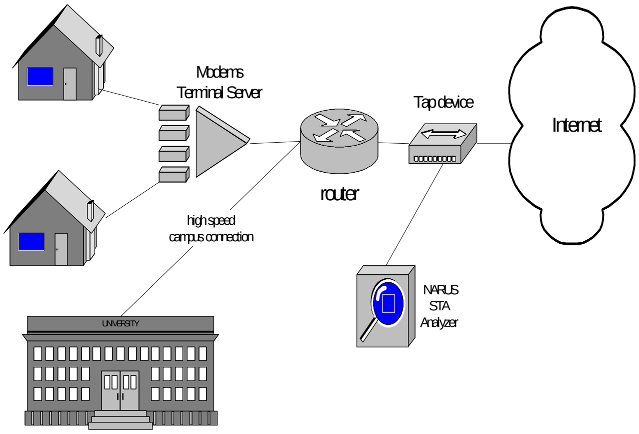

Figure 1: The placement of the Narus STA Analyzer.

Our collection, at a tier-2 ISP, started on Wednesday, November 3, 1999, at 10:53am local time and lasted 23 hours. There was to our best knowledge no significant international, local, sports or other event during this period of time.

The information collected contains over 3 million HTTP requests made from almost 20,000 (19,487) unique client IP addresses. Taking into consideration the high reuse rate for the dial-up IP address pool it is safe to presume that our collected data represents activity of more than 20,000 individual users.

We will only consider requests that result in actual HTTP object delivery (response status less than 300) where the object size is advertised by the server (“Content-Length” attribute is present). We have almost 2 million (1,925,075) such transactions in our collected data. Comparison of the object size advertised by the server and the actual size of the delivered object gives us a simple metric to distinguish cancelled and non-cancelled objects. Our data set contains almost 70,000 (68,481) canceled objects, representing 3.6% of all requests and we find that only 40% of previously canceled objects are later requested again by the same user.

Our analysis is done on a HTTP object basis where all HTTP transactions (requests) are considered to be independent. We unfortunately found it difficult to make any reasonable analysis on a HTTP document basis. This is not a limitation of the Narus STA Analyzer but rather a consequence of our experiment not collecting enough information to reconstruct the web graph. We are considering repeating our experiment over a longer period of time and also gathering information that would allow analysis on per-document basis.

C. Results

We first analyze the relationship between the QoE and the Response Time (aka Time-to-First-Byte) and then in the following section discuss the influence of the delivery bandwidth on QoE. All the graphs are based on the true object size, not including the HTTP and TCP protocol overheads.

1st. Relationship between QoE and Response Time

This section analyzes the time between the HTTP request and the first packet of the HTTP response. This time is a composition of the network latency (between the collection point and the HTTP server) and server processing time. We are unable to measure latency from the end-user host to the ISP network where we collect the information, but as a result of similarity in end-point equipment and small number of hops to the ISP network we assume this latency to be negligible.

Our data shows that over 95% of the requests had response time less than 500mSec and over 40% had response time less than or equal to 50mSec. Approximately 51% of all requests had response time in the 50-250mSec range. Estimation of the response time mean is 231mSec and median is 80mSec, which is consistent with previously reported results for roundtrip time analysis [10].

Figure 2 shows the cancellation rate as a function of the response time (network latency + server processing time) on a per object basis. Every point of this graph represents bins with 77,000 requests. Cancellation rate is calculated as number of canceled requests divided by the total number of requests in the bin. Average cancellation rate for our entire sampling is 3.6% (68,481 canceled requests from a total of 1,925,075).

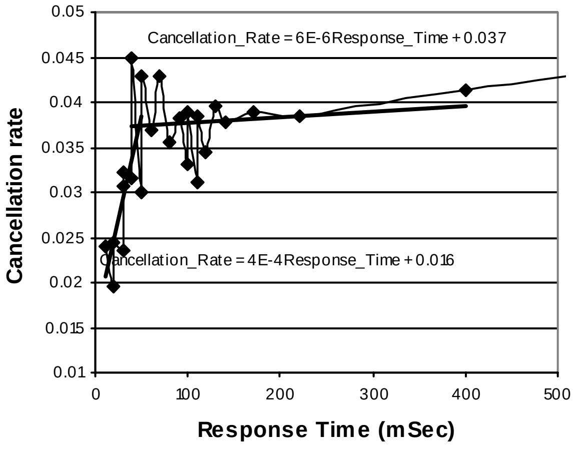

Figure 2: Cancellation Rate as a function of Response Time (network latency + server processing time, aka Time-to-First-Byte) with each bin containing 77,000 requests.

The effect of the response time on the cancellation decision is seen in Figure 2 to be negligible over the 50-500mSec range. The cancellation rate only increases from 2.3% to 3.8% as response time grows from 10mSec to 50mSec and then oscillates around 3.7% in the 50-500mSec range. We also have that cancellation rate for requests above 500mSec (5% of our sampling) is 5.5% which is significantly larger than the average over the entire dataset (3.6%).

Based on this analysis, we conclude that any additional efforts to improve response time in the 50-500mSec range will not result in significantly better user experience. It is important to note that majority of the Service Level Agreements (SLA) commit to provide network latency in the ranges of 100-300mSec. There is in our opinion no real difference in the level of service the end user receives for this range.

We find that additional improvement of the response time below 50mSec will result in better end-user experience. However, this is a difficult task at best considering that the round-trip-time (RTT) for a light beam traveling back and forth between the East and West coasts of the US is ~24mSec (and practical inter-coast network RTT is 100mSec or higher).

We have not analyzed samples with large (500mSec+) response time in this section. We assume that such large delays are related to the HTTP server rather than being caused by network latency. Though, this assumption needs to be investigated further.

2nd. Relationship between QoE and Effective Bandwidth

We saw in the previous section how QoE and response time are not strongly correlated. In this section we analyze how the effective bandwidth impacts the user satisfaction.

To get an accurate estimate of the delivery bandwidth, only objects for which 8kB or more was transferred have been included in the graphs shown in Figure 3 to 5. This leaves us with 373,050 object requests of which 22,903 (6.1%) were canceled.

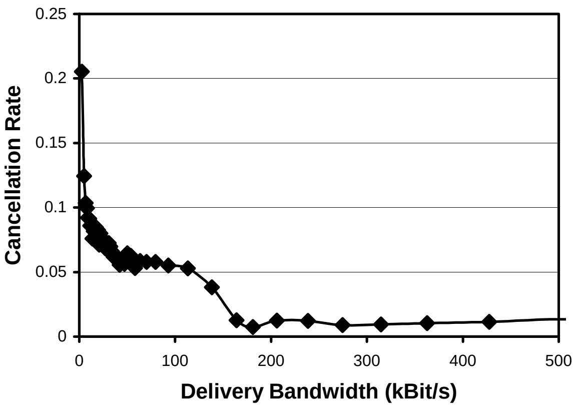

Figure 3: Cancellation Rate as a function of Delivery Bandwidth.

Figure 3 shows how the cancellation rate depends on the delivery bandwidth. Every point represents a bin of 7461 objects with a similar delivery bandwidth. Note that our ISP has two different types of customers: one group having dialup connections (up to 56kBit/s) and another group with a high-speed campus connection.

Figure 4 shows in more detail the low range of delivery bandwidth from Figure 3.

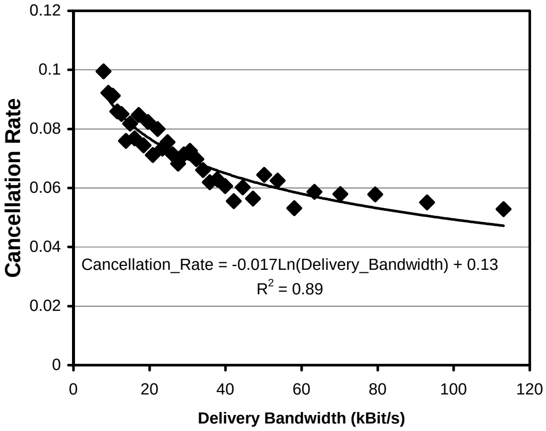

Figure 4: Zoom of Figure 3 for low range Delivery Bandwidth with a natural logarithmic best fit, the Delivery_Bandwidth for the equation shown measured in kBit/s.

We find from Figure 3 and 4 a strong dependency: the cancellation rate decreases sharply with improved delivery bandwidth. We find for objects 8kB+ that the cancellation rate is ~8% when delivered over a 14kBit/s connection, ~6% over a 56kBit/s connection and only 1% when delivered over a 200kBit/s connection. It is also interesting to note that we do not find any improvement in the cancellation rate for objects delivered over connections that are faster than 200kBit/s. Based on this analysis we conclude that effective network bandwidth plays a crucial role in end-user satisfaction. This means that upgrading a dial-up connection to basic cable/DSL results in dramatic improvement (cancellations reduced by a factor 6) of the user experience. However, we find no significant difference in customer satisfaction between basic (up to 200kBit/s) and high-speed (above 200kBit/s) connection services. Such precise data, instead of the obvious but vague 'faster-is-better' statement, can help cable providers in planning and deploying their services.

The curve fit in Figure 4 gives us the following estimate for the cancellation rate in the range 15kBit/s to 100kBit/s delivery bandwidth:

$$\text{Cancellation\_Rate} = -0.017\ \text{Log}_e(\text{Delivery\_Bandwidth}) + 0.13$$

3rd. Relationship between QoE and Object Delivery Time

Let us now try to express the relationship between delivery bandwidth and cancellation rate in a slightly different manner by relating cancellation rate to delivery time. For objects that were delivered completely our data immediately gives us the delivery time (Time-to-Last-Byte). But we have no delivery time for objects that were only partially transferred. So, for objects that were only partially delivered, the size of the object and the delivery speed up to the point where the user cancelled the transfer is used to estimate the delivery time. Figure 5 shows the combination of these actual and estimated delivery times.

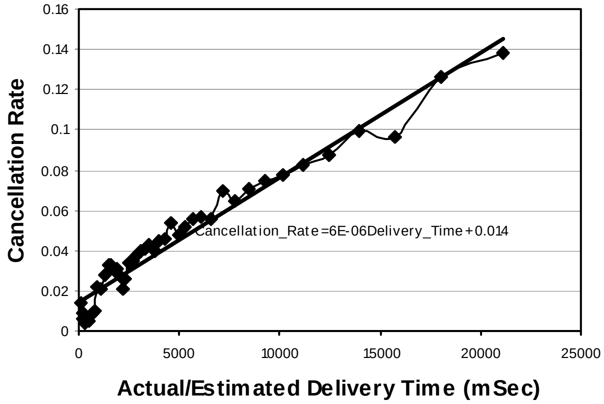

Figure 5: Cancellation Rate as a function of the Actual/Estimated Delivery Time. Average: 6900mSec. Median: 3400mSec. Bin size: 7461 objects.

Note first that a large portion of the samples are found in the lower end (the median is only 3.4s). This is probably caused by the object from the previous HTTP request being cancelled, where the cancellation request does not reach the server before the next object has begun transferring.

We see a nearly linear relationship for delivery time above 1 second:

Cancellation_Rate = 0.6% *
Delivery_Time_in_Seconds + 1.4%

This means that a 10 second increase in delivery time increases the cancellation rate by 6%.

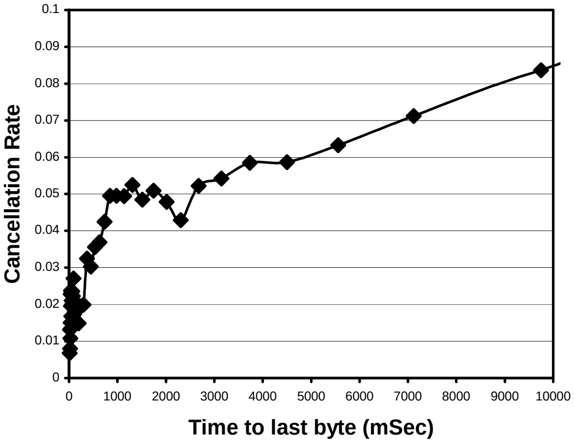

Figure 6: Cancellation Rate as a function of Time-to-Last-Byte. Average: 2500mSec. Median: 180mSec (this graph is based on the entire dataset without a lower 8kB limit as in Figures 3 to 5).

Figure 6 shows Time-to-Last-Byte using the complete dataset (we do not need to estimate the delivery bandwidth and do therefore not need any lower 8kB limit as in Figures 3 to 5). The 'last byte' event is either caused by the entire object being transferred to the client or by the user canceling the operation.

The data again shows lots of small transfers (the median is only 0.18s). We have that 90% of the objects have Time-to-Last-Byte under 8s.

Figure 6 shows less than 20% variation for the cancellation rate in the 1s to 3s range and a near-linear relationship above 2s. We furthermore see no sudden increase in cancellation rate around 8 seconds, which is opposite to the so called '8 seconds rule' [11,12,13].

D. Future Directions

We are planning to setup similar experiments for cable/DSL modem and wireless networks in order to compare user quality of experience over different types of networks. We are also considering the possibility of making similar analysis for international web traffic. An additional interesting challenge is to restore the object/document relationships and perform analysis on a per-document basis.

Interesting work could also be done by correlating web usage information (collected with the STA Analyzer) and account registration information (from RADIUS or DHCP servers). This would make it possible to analyze individual user behavior/actions as a function of the QoS of the network.

E. Conclusions

Our analysis based on objective metrics for web user satisfaction establishes a non-linear relationship between QoS and QoE. Based on this analysis we conclude that effective network bandwidth plays a crucial role in end-user satisfaction and that there is no gain in web browsing satisfaction for connection speeds above 200kBit/s.

We also find that network latency plays a less significant role on the level of user satisfaction, especially in the range of 50-500mSec where its influence is negligible.

Our empirical data challenges the validity of the current web performance ratings and the importance of network latency to the day-to-day user experience. Customers that do not use specific business applications requiring low network latency may choose not to pay high fees for Service Level Agreements (SLA) that guarantee such lower latencies since such provisioning will not significantly improve their user experience. Customers should/will instead invest in higher bandwidth network connections as these will have a much more significant effect on their day-to-day experience. Cable providers, on the other hand, can avoid over-provisioning their services by using the data presented in this article.

We also question the importance of the inter-server delay minimization as it is marketed by some vendors and by other "web acceleration" equipment manufacturers. We rather recommend that more development efforts be spend on improving bandwidth efficiency of the TCP stacks and HTTP servers.

In conclusion we advocate the necessity of more detailed investigation into the actual end-user QoE. The relationship between QoE and QoS is non-trivial and investigations into QoE could result in crucial tools for the planning and deployment of network based business and paid services.

F. REFERENCES

[1] ITU-T Recommendation P.861 - Objective quality Measurement of Telephone Band (300-3400 Hz) Speech Codec.

[2] "Managing Voice Quality with Cisco Voice Manager (CVM) and Telemate," Cisco Tech Notes, http://www.cisco.com/warp/public/788/AVVID/cvmtelemate.html.

[3] T. Berners-Lee, R. T. Fielding and H. F. Nielsen, "Hypertext transfer protocol - http/1.0," Informational RFC 1945, May 1996.

[4] R. Fielding, J. Gettys, J. Mogul, H. F. Nielsen and T. Berners-Lee, "Hypertext transfer protocol - http/1.1," RFC 2068, January 1997.

[5] A. Martin and J. Tai, "Workload Characterization of the 1998 World Cup Web Site", HP Lab Technical report , HPL-1999-35R1, http://www.hpl.hp.com/techreports/1999/HPL-1999-35R1.html, 1999.

[6] I. Marshall and C. Roadknight, "Linking cache performance to user behaviors," Computer Networks and ISDN systems, 30, pp.2123-2130, 1998.

[7] C. R. Cunha, M. E. Crovella and A. Bestavros, "Characteristics of www client based traces," Boston University Technical Report, BU-CS-95-010, July 1995.

[8] S. Waldbusser , "Remote Network Monitoring Management Information Base Version 2 using SMIv2," IETF RFC2021, January 1997.

[9] "NetFlow Services and Applications," Cisco White Paper, http://www.cisco.com/warp/public/cc/pd/iosw/ioft/neflct/tech/napps_wp.htm.

[10] M. E. Crovella and R. L. Carter, "Dynamic server selection in the internet," Proceeding of HPCS'95, August 1995.

[11] The Web Quality of Experience Workgroup, "The Web Quality of Experience Solution Architecture," White Paper, http://www.webqoe.org.

[12] Nicky Maraganore and Andrew Shepard, "Driving Traffic to your Web Site," Forrester Research, Inc, http://www.forrester.com, January 1999.

[13] Jacob Nielsen, "Designing Web Usability: The Practice of Simplicity," New Riders Publishing, 2000.