Integrating User-Perceived Quality into Web Server Design

Nina Bhatti, Anna Bouch[1], Allan Kuchinsky
Hewlett-Packard Laboratories, 1501 Page Mill Road, Palo Alto, CA 94304, USA
{nina, kuchinsk}@hpl.hp.com

Abstract

As the number of Web users and the diversity of Web applications continues to explode, Web Quality of Service (QoS) is an increasingly critical issue in the domain of e-Commerce. This paper presents experiments designed to estimate users' tolerance of QoS in the context of e-commerce. In addition to objective measures, we discuss contextual factors that influence these thresholds and show how users' conceptual models of Web tasks affect their expectations. We then show how user thresholds of tolerance can be taken into account when designing Web servers. This integration of user requirements for QoS into systems design is ultimately of benefit to all stakeholders in the design of Internet services.

Keywords: Quality of Service, User perception, E-Commerce, Server design

1. Introduction

The success of any scheme that attempts to deliver desirable levels of Quality of Server (QoS) for the future Internet must be based, not only on technology improvements, but on users' requirements [15,19]. To date, the majority of research on QoS is systems oriented, focusing on the scheduling and routing of network traffic. Although it is often recognized that a measurement of user satisfaction must be included in assessing network efficiency [20], relatively minor attention has been paid to user-level QoS issues. The number of Electronic Commerce (e-commerce) users is rising. As the e-commerce industry grows the topic of providing adequate QoS for the Internet becomes increasing critical to businesses. To provide the flexibility needed to respond to customer requests, Web pages that support e-commerce applications typically are dynamically computed. This means that the delays witnessed by users are directly affected by server performance, and not simply due to download times. Inevitably, more requests are made of servers than they can immediately handle -- the magnitude of user demand outstrips server capacity. The outcome of this situation is that often some users are denied access to the server, or the accessed service is unacceptably slow.

As the World Wide Web is rapidly increasing with numbers of users expected to reach 320 million by 2002 [35], the increase in network usage is paralleled by a growing diversity in the range of supportable applications. Because of its accessibility, the future Internet offers the potential to break traditional barriers in communications and commerce. However, the current service to users is often unacceptable [12,36] and is likely to remain so in at least the near future [5].

The components of QoS systems are extremely difficult to integrate. For example, server utilization cannot be divorced from the requests made to that server from applications, or from network conditions. For example, providing another 5% worth of server utilization may require a considerable amount of computational effort, but have minimal incremental benefit to users. A difficult but central question for server designers in the future is to what degree user perception of improved quality of service can be translated into metrics that can be used to inform service providers in designing resource allocation strategies. The real challenge for future network designers, therefore, may not solely lie in maximizing utilization of servers, but in ensuring that the service provided is both efficient and subjectively valuable to users [7].

This paper reports results from a set of studies into how users define and perceive Internet QoS. We describe empirical work that shows that a mapping can be developed between objective and subjective expressions of latency. Latency is defined as the delay between a request for a Web page and receiving that page in its entirety. We chose to study latency, not simply because it is associated with the most common cause of poor QoS, but because it represents a problem that is likely to escalate as Internet usage inevitably grows [22]. We use qualitative data to elucidate the motivations behind behavior observed in empirical work. We then show how these results can be included in server design to improve realized QoS. Our server designs use prioritization schemes that attempt to meet

the increasing demand for access to network bandwidth and server resources according to the QoS needs of applications [4,14,16,30]. Priority scheduling schemes can be implemented in the server using mechanisms that queue and service traffic from particular applications in a specific order. Schemes such as differentiated services exploit this ability by classifying packets of information in certain service profiles [4]. It should not be assumed that the requirements of applications regarding QoS can be divorced from the requirements of those who ultimately use those applications [30]. However, it was not known to what extent objective QoS metrics relate to user perceptions of quality and impact the behavior of users. Only by understanding this relationship can we define the potential trade-off between the cost of resource allocation for the service provider, and the benefits in increased business gained by providing a level of QoS perceived as valuable by users.

The rest of this paper is organized as follows. Section 2 describes the objectives and design of experiments to assess user's tolerance for delay in an e-commerce setting. Section 3 describes the results of these studies. Section 4 shows how these results can be incorporated into server designs to achieve improved perceived QoS for users. Finally, section 5 wraps up with some remarks about future work and our conclusions.

2. Experimental assessment of users tolerance for delay

2.1 Objectives

Measured thresholds for QoS are increasingly important to system designers [17]. Establishing a mapping between objective and subjective QoS is perhaps the most direct way research can enable servers to be designed to provide maximum utility. A common finding from previous research is that QoS delivered by servers must accord with users' expectations in order to be perceived as acceptable [7,8,33]. Objective measurements, such as response time and delay cannot, however, fully characterize the factors that drive these expectations. A consistent finding is that QoS received by users should concur with their expectations but that these expectations change according to the pattern of quality received [8,10]. The characterization of factors that impact users expectations is complicated by the fact that many such factors are interrelated. For example, Web pages that are retrieved faster are judged to be significantly more interesting than their slower counterparts [28], and users may judge a relatively fast service to be unacceptable unless it is also predictable, visually appealing and reliable [8]. Indeed, the weight of evidence from prior research suggests that there is no direct correlation between objective levels of quality received by users and their perceptions of that quality. To predict users' tolerance for QoS it is therefore necessary to understand what motivates users' judgements of QoS. We selected study participants that met the profile for e-commerce users that were moderately experienced Internet users. The measures of QoS were established during a representative set of e-commerce tasks. We set out to define the minimum latency users would tolerate before they find that level of QoS unacceptable and potentially take their business to a competitor.

Our study enabled us to address the following questions:

- Are there objective measures for user's tolerance of delay?
- Is this tolerance affected by the task?
- Is this tolerance affected by the duration of interaction with the site?
- What is the perception of businesses that have poor QoS?
- Does web site design influence user's tolerances?

2.2 Experimental Design

We set out to design experiments that would allow us to assess whether we can measure a user's tolerance for delay and what affects this tolerance. To gather this information we created a Web site with delay and bandwidth programmability. This provided a self-contained and consistent Web shopping experience.

Because we focused our interest in Web QoS for e-commerce, we selected participants that would match a profile of Internet shoppers. There were 30 male participants, between the ages of 18 and 68 [10], in the study. It was essential that a homogenous group of users was selected, because users with different amounts of knowledge and experience of Web QoS have different expectations of QoS [8]. We restricted our sample to participants who:

- Use the Internet for at least 2 hours per week
- Have made at least 2 purchases on the Internet in the last year.
- Have at least an intermediate level of self-assessed skill with using computers.

Male participants were selected for the study to eliminate any confounding effects due to gender differences in visual perception and learning [23]. Males were identified as the most frequent users of Internet services [27].

The participants were given the same task series so that their path through the site would be consistent. The task involved configuring and purchasing a home computer system. We wanted the task to be as ecologically valid as possible. We therefore chose an e-commerce site currently in operation, the HP Shopping Village [34], www.shopping.hp.com. This busy site is ranked first for retail revenue generated by e-commerce[17]. This situated the task strongly in a real-world context. During the task participants were asked to purchase each component of the computer separately. To answer the research questions, it was necessary that the chosen task meet certain criteria. To study whether users' requirements for QoS were similar for similar sub-tasks, a set pattern of actions was repeated through the task. For each component purchased, participants were required to:

1. View a class of similar products.
2. Select a specific product from a class of products.
3. Add the chosen product to their shopping cart.
4. View the contents of their shopping cart.

We needed a consistent set of tasks to determine if users' tolerance changes over time. If the tasks had been widely variable, then any change in tolerance could be ascribed to the variation in what participants were asked to do, and not to a genuine accumulation of frustration. The task was designed so that all participants followed the same path through the Web site.

Participants gave feedback on Web performance satisfaction through:

- interaction with a quality rating browser extension
- verbal protocols -- participants talked aloud while performing the experimental tasks
- participation in focus group discussions

We correlated user feedback with actual known delay measurements and built a model of users' tolerance. The capture of QoS acceptability information, in our study, was driven by the correlation of user interface button clicks for rating of QoS, and verbal protocols. The inclusion of verbal protocols enabled us to gather feedback from users during interaction with the web site. Focus group studies enabled us to explore in a wider context the issues raised during the protocols.

2.3 Experimental conditions

The first condition in the study investigated how the latency between requesting a page and receiving it is perceived by users. Varying the latency in this condition has the effect that the page where the link has been clicked remains displayed in the browser until the next page has been loaded. This next page is then brought up in its entirety. Predetermined delays ranging from 2 to 73 seconds were injected into the loading process. The choice of this range of speeds was guided by speeds that users had perceived to be qualitatively different in previous research [13,28]. There were two sequences of delay for latency. Pattern 1 mimicked a random pattern of delay. In pattern 2, the delay generated on the Web pages was relatively smooth.

Experiment 1 Classification of latency

Participants were asked to perform the shopping task and rate the latency received for each Web page access. An interface was developed to register ratings. The inferface contained grey buttons labeled "high", "medium", "low", as well as a black button, marked with an "X". Participants were directed to click one of the buttons in this interface for each Web page accessed. Participants were told that the black button, marked with an "X", should be used to indicate that the quality was totally unacceptable.

Experiment 2 Control of latency

In this condition, users were told that if they found the delay of the Web page unacceptable they could click a button labeled "Increase Quality". The effect of this button was to immediately bring up the requested page. Previous research suggested that this would be a valid measure of users' requirements for speed [6]. This experimental set-up contrasts users' opinions about tolerance of QoS, captured in classification conditions, with what can be implied about users' tolerance from their behavior when they controlled the quality.

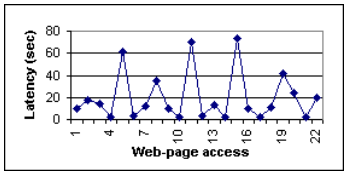

Figure 1: Incremental Loading Latency

Experiment 3 Incremental Loading

This part of the study investigated whether users would be more tolerant of delay when Web pages loaded incrementally instead of all at once. Previous research suggests that providing continuous feedback reassures users that the system is working and gives them something to look at while waiting [24]. However, [26] points out that standard browser feedback, provided in the form of progress bars, fails to communicate the amount of the page that has been completed. Loading Web pages incrementally can address this shortcoming, while providing users with visually interesting feedback. The flow of information between Web server and client was manipulated to cause the Web pages to load in parts. In our task, participants would receive the banner of the next page as soon as they clicked a link. This was followed by text, and later, graphics. Participants were asked to evaluate the time it took for the whole Web page to complete using the same GUI as in experiment 1. The time taken for specific pages to complete was random. Figure 1 shows the mean delay taken by each Web page to complete in condition 3. These measurements were taken using client-based software that captures latency received by users with 100% accuracy [32].

Participants were split into two groups for the investigation of latency in experiments 1 and 2. 15 participants received pattern 1 for both the classification and control of latency, while the remaining 15 participants received pattern 2 for both the classifications and control of latency. Table 1 summarizes the experimental conditions applied.

Experiment	Pattern	Participants
1 (classify)	Random	15 (Group 1)
2 (control)	Random	15 (Group 1)
1 (classify)	Regular	15 (Group 2)
2 (control)	Regular	15 (Group 2)
3 (classify)	Random & Regular	30

Table 1: Experimental Conditions

3. Results

The key finding of the research was a mapping between objective QoS and users' subjective perceptions of QoS. The data we gained from verbal protocols and focus groups indicates that participants were strongly influenced by their expectations of the delay when responding to the QoS they received in the experiment. Additional discussion of experimental design and results can be found in [9].

Focus group data indicated that tolerance of delay was decreased when there was a conflict between the level of quality expected and that received. We found that there was almost unanimous agreement among participants concerning the factors that help form these expectations. These expectations are influenced by contextual factors including the type of task, the method of page loading, and cumulative time of interaction. We also found that there are very real business consequences of slow server response times. Users believe that if performance is poor, the security of the site may be compromised. Poor performance also leads to loss of customers.

3.1 Measures of users' tolerance for delay.

Verbal protocols indicated that participants used the "Low" button when they found the QoS was unacceptable; very few participants used the black button labeled "X". We took this into account by aggregating the "Low" button and "X" button responses when conducting a set of Chi-squared tests for statistical significance. Table 2 lists the ratings of specific delays and shows that the threshold where QoS is judged as "Low" is around 11 seconds. This finding is consistent with previous work that established this threshold for holding users attention to the task [11]. The range of latency assigned by participants to each classification in condition 3 (incremental loading) are almost 6 times higher in each case compared to the classifications made in condition 1 (see Table 2).This indicates that users are more

tolerant of latency when Web pages load incrementally than when there is a delay followed by the display of the page in its entirety. These results indicate that incremental loading may help to maintain users' attention to the task at hand, rather than to the QoS they receive. Furthermore, Table 2 shows us that, relative to the selection of "Low" or "Average", quality of service is more likely to be classified as "High" in condition 3 (incremental loading); this category is proportionately much larger in experiment 3 compared to experiment 1 (classification of latency).

We observed, in experiment 2 (control of latency) that there was a large standard deviation among participants in terms of their tolerance of latency, Although the average tolerance was 8.57 seconds in this experiment, the standard deviation was 5.85 seconds. It is not possible for us to conclude from experiment 2 that users will tolerate a specific amount of latency before finding that QoS unacceptable. Multiple regression analysis revealed that the number of hours participants used the Web significantly influenced their tolerance for latency. Higher levels of Web usage were associated with less tolerance for delay during interaction ($p<0.01$). The large standard deviation observed when participants were asked to control latency may have been due to the differences among participants in terms of their risk-taking behavior. Participants differed in terms of whether they took advantage of the fact that there was no penalty for pushing the button to increase quality. This difference is also suggested by the fact that there was no correlation between participants' tolerance when classifying latency and their tolerance when controlling latency. To gain useful insights from this condition, we therefore investigated the levels of tolerance demonstrated by each individual participant.

Rating	Range of Latency experiments 1 & 2 (non-incremental loading)	Range of Latency experiment 3 (incremental loading)
High	0 - 5 sec	0 - 39 sec
Average	> 5 sec	> 39 sec
Low	> 11 sec	> 56 sec

Table 2: Rating of latency

Previous research has established 3 thresholds relating to users' tolerance for delay [26]. For delays of 0.1 seconds or less, users perceive the response as immediate. A delay of 1 second corresponds to the pace of an interactive dialog. A threshold of 10 seconds was identified as the point at which a significant number of users perceive the delay to be unacceptable. According to this research, a 10 second delay corresponds to threshold where users lose their attention to the task at hand. These findings fit with the literature on cognition [11]. Two stimuli within 0.1 second of each other fuse into a single precept, e.g. two pictures seen within 0.1 second fuse into a perception of motion; animation breaks down if longer than 0.1 sec/frame. The coarsest level of interaction is the "unit task", the pace of routine cognitive skill, e.g. 10 seconds is about the time needed to select text on a screen and modify it. Thus, a delay of over 10 seconds constitutes a disruption in the "unit task" and may cause disorientation and reduced performance. Other research has described thresholds that are perceived as qualitatively different over a wider range of latency [28].

3.2 The duration of time users interact with the site

Investigating whether tolerance for delay is influenced by length of session is especially pertinent in the area of e-commerce. On one hand, users' frustration at delays incurred may accumulate. This would mean that they would tolerate less delay as the session time increases. This is a likely scenario because it has been shown that users conceptualize the quality of their interaction according to their ability to reach the top-level goal. In the case of e-commerce this goal often is to make a purchase. Furthermore, in e-commerce, subtasks are by nature structured so that the act of purchasing is normally the last in a chain of related operations. If users' tolerance for delay decreases over time, then this has clear impact for loss of business on the site. On the other hand, as the length of the session increases, users have invested more time towards reaching their goal of purchasing a product and thus have an incentive to continue. It may be that, as the time remaining to complete their task decreases, users' tolerance for delay goes up.

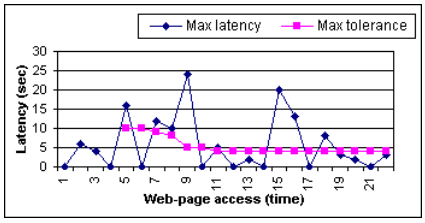

Figure 2: Tolerance of Latency over a session

The first condition in the study investigated how the latency between requesting a page and receiving it is perceived. In all conditions we found that users' tolerance for delay decreased as the length of time they spent interacting with the site increased. In all cases this finding is statistically significant ($p<0.01$). The effect is more powerfully significant for condition 3 ($p<0.001$). Figure 2 is an example of the maximum delay tolerated by a participant in condition 2. Maximum tolerance for delay is represented by the point at which the participant clicked the "Increase Quality" button.

Our results suggest that users become increasingly frustrated with delays incurred during interaction. Qualitative data shows that although users are less likely to leave an on-line shopping site once they have placed objects in their shopping cart, they are no more tolerant of delay. In fact, users are more likely to become annoyed in this situation as they feel they have less control over interaction and have been manipulated into being forced to endure poor QoS:

`I'm already half way through what I wanted to do, now I'm caught because I can't leave, but I won't come back'[2].

3.3 Expectations based on task.

The user's goal, when interacting with any network application, has been shown to affect not only the level of QoS that the user will tolerate but the very definition of quality [30]. For example, requirements for high video performance are more prominent in interactive tele-teaching tasks than in listening to lectures [21]. Furthermore, it has been established that large quality variations should be avoided for audio transmission [31]. We set out to investigate the influence of users' tasks on their tolerance of delay in the e-commerce environment. Our findings suggest that there is a distinction to be made between a situation where a user interacts with a Web site for information gathering purposes and where that user interacts to undertake a specific action (in our context, to buy an item):

`...it depends on the intent. If I'm browsing for something then the (quality) I get isn't so important as if I've got a definite mission in mind'.

The type of real-world task in which they are engaged is likely to be involved in forming expectations of QoS and therefore have an influence on the amount of delay tolerated:

`When I've added stuff to the list...I would expect it to take a little longer than when it's got preset pages'.

`Like when you're comparing I expect that to take a little longer because it's going to have to go out and get information'.

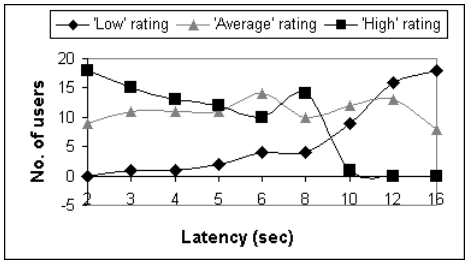

Figure 3: Rating of latency by users.

Qualitative data suggests that participants expect different tasks to take longer than others. From this information we were able to classify tasks according to participant's expectations of the latency each task should incur. High tolerance tasks were:

- Comparing several items.
- Viewing the shopping cart.

By comparison, low tolerance tasks were:

- Returning to a previously accessed page.
- Viewing a class of products.
- Adding to the shopping cart.

During the experiments we found that users tolerated different levels of latency depending on what they were doing. As can be seen from Figure 3, participants classified response with an 8 second delay (corresponding to comparing different printers) as higher quality than that with a 6 second delay (corresponding to viewing a class of monitors). Statistical tests show that users will accept more delay when they are comparing products or viewing the contents of their shopping cart than when they are viewing a class of products or adding to the shopping cart, ($p<0.01$).

Qualitative data showed that participants had a conceptual model of the way that networks store and access information. This conception influenced users' tolerance for delay. For example, our data indicated that tolerance for delay associated with specific tasks was dependent on *a)* if the user believed that the task required accessing a database, e.g. from which to compare products, *b)* if the task involved a calculation to be made, e.g. in calculating the total spent from the items placed in the shopping cart.

`when I brought up my shopping cart I figured it would have to compile a bit longer so I was more willing to wait a little bit for it to come up'.

In an e-commerce environment different tasks imply different levels of economic incentive on the part of the company whose products are sold. For example, participants expected tasks like adding to the shopping cart to be relatively fast because of the company's motivation to encourage user's to make a purchase. Although no pages were cached in the experiment, participants awareness of this technology made them relatively intolerant of delay when re-visiting previously accessed pages.

Our findings suggest that users anticipate the amount of time it will take them to perform particular on-line tasks. This anticipation helps form their expectations of the time it should take them to complete a whole task. Our results suggest that when the process of completing a task is disrupted by unanticipated delays, a conflict arises between users' expectations and the QoS they received. This conflict results in a rating of poor QoS:

`So I'll be sitting there for half an hour so I'm set for that...so a lot of it depends on the time I anticipated I had when I set out'.

`If I'm going to buy something that I need to do research on, mentally I'll allocate more time'.

3.4 Feedback

If delivered quality is to concur with users' expectations, that service must be predictable [25]. Previous research has established thresholds by which Web page response times can be classified, and related these thresholds to the need for browser feedback to enable users to predict response times [20]. If the delay from request to the display of a Web page in the browser is one second or less, then there is no interruption to users' flow of thought. Users perceive this response to be immediate and therefore do not require feedback in the browser. However, in the frequently occurring situation where servers cannot provide an immediate response, continuous feedback must be provided. Feedback is especially important if the delays incurred are likely to vary, as in [6]. Feedback enables users to predict the amount of time they will have to wait. We investigated the interaction between providing feedback to the browser and overall judgements of QoS by comparing users' tolerance for delay under two conditions:

1. Page loads incrementally: The Web page would be brought up in parts. In our study this meant that users would receive the banner heading the site first, followed by graphics and, later, textual information.
2. All information displayed together: The Web page from which users clicked the link would remain displayed in the browser until the entire next page could be downloaded.

Consistent with previous research [26], we found that, in circumstances where feedback is provided (condition 3) tolerance of delay is significantly higher.

Qualitative data suggests that the value of feedback is that it:

- Promotes confidence that users' requests are being processed:
  `As long as you see things coming up it's not nearly as bad as just sitting there waiting and again you don't know whether you're stuck'.
- Enables users to estimate how long they will have to wait until they can interact with the site:
  `Well I know if it's saying 33% or whatever then I'll have to wait a couple of seconds'.
- Focuses their attention by giving them something to look at while waiting:
  `at least you're not sitting there with nothing to look at, while I'm waiting for it to come up, I can be reading'.

Some participants in our study used the standard browser feedback – messages of percent download completed within a small status bar to assess activity in the network. Typically, these participants did not prefer incremental loading. This finding confirms the strength of incremental loading as being indicative of the processing of a request. Either browser feedback or incremental loading can provide this feedback.

3.5 Business Implications of Poor QoS

Recent assessments of Web usability indicate that the same QoS dimensions are responsible for the greatest number of degradations in users' perceptions of overall QoS for over three years [18]. Several prominent Internet sites such as www.ebay.com, www.schab.com and www.brittanica.com have all experienced publicly embarrassing unavailability and poor performance. In a review of twenty prominent sites it was found that what is called the greatest "design mistake", slow download times, was committed by an average of 84% of web-sites. This figure is likely to be even greater for the aggregate of Web sites, since smaller companies often provide lower levels of QoS than prominent companies. A recent study of nearly 3000 on-line shoppers found that people use e-commerce sites because they are convenient. If systems designers cannot understand the limits of users' tolerance for slow download times they risk not only promoting users' frustration but also an eventual and significant loss of business.

We found that users' perceptions of the QoS they receive effects not only the likelihood of going to a competitor's web-site but also their opinion of the company's products and of the company itself. A failure to understand users' on-line QoS requirements, therefore, may affect users' conception of a company's stature and commercial viability. As more and more people use the Internet for commerce, service providers must integrate users' QoS requirements into their server design in order to meet the needs of *their* customers, the retailers whose products are sold online.

Data from users suggest that blame for poor QoS is placed on the server, even though the users in our study possessed a conception of the manner in which data was routed on the network. Indeed, although participants said that they could reason that poor QoS could be due to the amount of traffic on the network, they nevertheless did not intuitively associate this situation as the cause of delay. When participants were questioned about the causes of delay they did not blame network traffic demands, networking infrastructure, ISPs or even their own modem connections; instead they placed the blame on the individual businesses represented by the sites.

3.5.1 User's expectations of corporate Web sites

Inevitably, if poor --or unpredictable-- QoS is habitually experienced at the site of a particular company, the products of that company are likely to be viewed as inferior. Participants in our study believed that companies that are more commercially successful should possess the financial means to supply at least adequate levels of QoS 100% of the time. This expectation means that users are less likely to accept delays, or refused admissions, to a site that promotes the products of high-status companies:

`Because the companies are so huge they should pour money into their web-sites, should have fast sites. If I try to get on those sites and they're slow then I'm not as patient'.

`This is the way the consumer sees the company...it should look good, it should be fast'.

Qualitative data also shows that users who frequently purchase products from particular web-sites habituate to the typical levels of quality they receive from those sites. Conceptually, this leads to a sense of betrayal if the QoS delivered is not according to what is expected. Users describe this situation as compromising their conception of the customer loyalty shown to them by the company from whom they are buying:

`If I've been going to you for a long time and you suddenly can't perform...well then you've sort of betrayed me and I won't be going back'.

Unpredictable service therefore compromises users' trust in the company. Failure to provide a consistent level of QoS that is needed to maintain users' sense of customer loyalty means that users will not return to that site.

Maintaining acceptable levels of QoS is not just the problem of the service provider. The companies whose products are sold on-line, and the advertisers who are their sponsors, are also affected by the ability of Web servers to provide acceptable QoS to users. Our data show that the ultimate consequence of falling short of this goal is loss of customers.

Users have too many Web sites that they can use as alternatives if they are either refused entry to one site or are given particularly slow service. There are almost no barriers to switching to another site if performance is unacceptable. This makes performance critical to attracting and keeping customers:

`There's just too many alternatives, I can't think of anything that I can't just go and get somewhere else'.

3.5.2 Compromised Security

Another finding in our focus groups was that users made a connection between poor performance and compromised security. Participants in this study felt that cumulative slowness on Web pages suggested, not only that the products being sold were of inferior quality, but that the security of their purchase was compromised:

`If it's slow I won't give my credit card number'.

`I'd say, you haven't got your resources figured out, you're a poorly managed outfit, I don't trust you any longer'.

Once users perceive security has been compromised, no purchase will be made and the main purpose of any commercial Web site becomes critically compromised. It is therefore crucial for systems designers to understand the effect of cumulative frustration, especially as it is typically in the later stages of interaction that users are likely to commit to a purchase.

3.5.3 Deferring users

There are inevitable spikes of traffic that can overwhelm a server and therefore admission control may be used occasionally. We asked participants about their opinions regarding being denied access to a site. Opinions were very negative:

"Could you imagine going to store and someone saying, `oh, too many people waiting in line, come back in an hour'."

"You're going to go to the next store."

Users' conception of the Internet is that it provides service on demand. Indeed, the success of the Internet is in part due to its convenience. Our evidence shows that users define *convenience* as "accessibility" and "ease of use". In the same way there is a conception that companies want to encourage visitation to their site. This is especially the case in the realm of Internet commerce. Asking users to defer their requests is therefore in direct opposition to users' concept of service on demand. We found that if sites must defer customers then some sort of incentive should be offered to return to the site.

Participants suggested that they would be more willing to defer their request if a discount or "coupon" was offered as an incentive to go back to the site:

`We'll give you 5% off if you come back, well, that would be a different story'.

3.6 The effects of Web site design

Data from the verbal protocols suggest that the actual design of a Web site can have a profound effect upon the perceived Quality of Service in several areas:

3.6.1 Page structure

Previous work has shown that users judge the speed of the service they receive according to their ability to accomplish the overall goal of their task [25]. Verbal protocols taken while participants performed the task indicates that the time taken to scroll down a specific page to locate a link detracts from users' perceptions of page latency. For example, participants reported annoyance with situations that required them to scroll a page to reach the desired information. They were particularly disconcerted in cases where they had to scroll in order to locate the selection to make for adding an item to the shopping cart. This is an example of an interaction affect between page structure and

overall perceptions of QoS:

`The speed wasn't bad except (when getting) the paper...plus the fact that you have to scroll down to view your basket'.

3.6.2 Iconic representation

The use of icons in interactive applications has the benefit that they are more easily associated with real-world metaphors than text-based information [29]. The functionality of a real-world metaphor can be encapsulated in a simple pictorial representation. The use of icons is, therefore, especially relevant when the intention is to associate the functions of a real-world metaphor with a well-known image. A prime candidate in our study would be the use of a shopping cart. Indeed, many participants suggested that this would have been an intuitive use of graphics, enabling them to clearly see the functionality of the site.

In our task users wanted an "Add to Shopping Cart" button be placed prominently on all pages of the Web site and be easily accessible throughout all phases of the shopping experience. This would certainly be in the interests of the proprietors of an Internet commerce sites, who wish users to add items to their shopping carts. The fact that this wasn't the case in our study often led users to associate the site with an overall lower standard of service.

3.6.3 Number of links

Another issue for Web site design, which effects perceived quality of service, is the number of link traversals necessary for a user to reach information of interest [18]. It should be noted that improvements to server and/or network performance can only improve the delivery of each individual Web page. If there is an excessive number of Web pages that have to be visited before the item of interest can be retrieved, then the benefits of server and/or network performance improvements will go unnoticed. This implies that the proprietors of electronic commerce sites should apply sound principles of information structuring -- e.g. link trees should be wide, rather than deep and that such information structuring decisions are as crucial as server performance is to the perception of quality of service.

The point to stress is that the quality of the Web site design is inextricably bound with the perception of quality of service and that any attempts to improve quality of service should include site design as one of the parameters to be considered. Inevitably, providing users with optimum levels of QoS involves sometimes subtle trade-offs between maximizing the ease of use of a site and the speed at which it can be delivered. An obvious example of this is providing iconic as opposed to text-based information. What our results show is that users' overall perceptions of the performance of a site is affected, not just by the objective latency of each page, but by the delay incurred during their interaction with that page. This latter delay can be due to poor Web site design.

4. Implications for server design

The perspective of this work is not only to understand user behavior relating to QoS, but to interpret those findings into solutions for real-world problems. Our findings have implications for the way that servers dynamically control the processing and delivery of information in response to users' requests. For example, Web servers can be altered to modify the scheduling of requests so requests are served more selectively than with the traditional FIFO mechanism. [1,2,6] give architectures for modification of Web servers to allow control of scheduling of requests and resources given to these requests. There are also several operating systems efforts underway to account for and control system resources given to each class of web request [2,3]. While we have the technology to better control the level of service each web request receives, little work has been done to define and implement policies based on user perceptions of quality of service. In this section, we provide some insights into appropriate policies for web server QoS controls that can adjust the server response time to more closely match the expectations of users, therefore maximizing the utility of the sever.

4.1 Meeting latency requirements

To facilitate user satisfaction all requests should be processed within the latency requirements for high QoS rating given in Table 2. We propose to modify the scheduling algorithms of Web servers to ensure that tasks complete within their deadlines. This can be accomplished, for example, by Earliest Deadline First scheduling. Each request that enters the web server has an associated deadline for completion. The association of deadlines with requests can be fairly fine-grained; the deadline associated with the request can be based on the task. From the reslults of our study it was clear that users have different models in their mind about which tasks "should" take a while and which tasks

"should" be fairly quick. These different deadlines can be assigned by parsing the URL for the request, where the URL has been encoded to indicate task urgency (see Section 4.3), and then associating the correct deadline. The server is given fine-grained information about which tasks are expected to finish quickly.

Current Web servers such as Apache are a collection of processes and execution threads that all implement a fairly simple model: wait for a new request, accept the request, process the request, send the results to the client. There is no control over which request completes first. Requests are simply executed as soon as a server process/thread is available and can accept the connection. This means that a potentially important request with very low latency requirements can be waiting in the queue with no process assigned to service it. By the time a process accepts the request and determines its scheduling precedence, the deadline may have passed. This is an especially critical problem when servers are busy and therefore the delay increases before a process is available to handle the request. Web servers can be modified to accept all connections quickly for classification, then, after consulting QoS policies, calculate deadlines for each request. The server processes then can select the next *most urgent* request to complete.

One of the most efficient ways to improve server resources is to complete work that has higher value. We can use our findings of objective thresholds to enable servers to process requests while they still have utility to users. It does not benefit users if server resources are wasted on requests that have been waiting so long that it is likely that the user has long since moved on to other web pages. The ability to associate timeliness data with each request allows the server to be more selective in its scheduling

In addition to providing better service to all users, our objective measures can associate target performance for different classes of service. Current proposals for differentiated QoS [2,6] are driven by relative measures of performance, e.g. in best-effort vs. premium service. They are not based upon absolute measures of performance.We can only assure that a premium client was receiving better service than a best-effort class of service client. When a premium client receives slow responses it is not reassuring to tell the client that there are users who have lower priority. Using our data for high, average, and low response time ratings we can associate specific deadlines with different classes of service. Each request can be marked with not only the priority of the request but also with a specific deadline that satisfies the specific class of service. This allows the server to offer differentiated services, based not on relative priority scheduling, but upon actual performance within defined objective measures.

In the case of differentiated services where a server provides OS-level support for allocation of resources based on class of service, the performance of each class can be compared to what the targets are for the class. If the target performance is not attained, then the server will have to allocate more resources to the premium class. This is an important auditing and control mechanism. Otherwise we know only that a class of service has been given a certain share of say CPU resources; it may still be failing to achieve the performance goals.

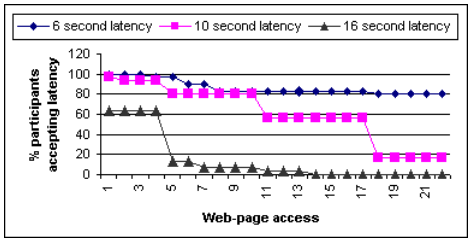

Figure 4: Tolerance for latency over time

4.2 Duration and latency requirements

A central finding in our study is that users' tolerance for latency decreases over the duration of interaction with a site. Figure 4 shows that this affect is apparent for both relatively low and relatively high levels of delay. A 16 second latency is acceptable to 60% of the participants during the first 4 web page accesses, but not acceptable to anyone for accesses over the 13th page. This is extremely significant, as e-commerce sites often have a fairly complex organization where a transaction is composed of many web page accesses. A 6 second latency was rated as acceptable for all participants until the 3rd page access and then the number of users that rated it as acceptable declines steadily to 80% for 20 or more accesses.

If an e-commerce site wishes its customers to rate their shopping experience as acceptable, then the site must

assure that the performance does not degrade. This can actually mean that the latency must improve over the duration of a session. This has a particularly profound effect upon the maintaining of ongoing customer relationships, i.e. ensuring repeat business. Our focus groups found that users will complete a shopping transaction even while the site has poor performance, if only because they have loaded up their shopping cart and therefore have a significant investment of time and energy at the site. However, if performance is perceived as not acceptable, the customer will remember that shopping experience and actively avoid the specific site again. This results in perhaps one completed sale but more importantly one customer who will not come back to the site. This information about a repulsed customer is not contained in any log; only a successful transaction is recorded, providing misleading information to the e-commerce site operator. To ensure repeat business, the transaction must complete with acceptable performance for each page access. The phenomenon of duration causing users to be more critical of performance can be expressed as a function that takes into account the duration of the session. Here we express the *utility* of a session of requests as a number between 1 and -1, where *utility* > 0 indicates that acceptable performance thresholds have been exceeded, *utility* = 0 indicates that the service is exactly acceptable, *utility* <0 is unacceptable performance. Total utility of as session of length *N* can be expressed as:

$$Utility = \frac{\sum_{i=1}^{N} \frac{threshold(i) - latency(i)}{threshold(i)}}{N}$$

threshold(i) = threshold of acceptability of access *i*
latency(i) = delay in completion of page *i*
N = length of session

Based on our data, *threshold(i)* decreases as *i* increases which implies that *latency(i)* must improve as *i* increases, just to maintain the same utility over a session. This relationship is illustrated in Table 3 below.

An e-commerce site generally has a notion of a session which is typically implemented using cookies. A cookie is used to associate a user with a shopping cart and profile. The cookie can directly encode a session duration field and be used to index into a table of session duration. With this information the server can schedule response times to maximize the utility for each user, for the average of all users, or for a premium class of users. The server can use the session length as an indication of the tolerance of the user to delay. At a minimum, this calculation of utility can be used to more exactly assess the performance of the site by taking into account the session duration. An alternative to this cookie method is the use of URL parameters that encode the length of the session. This is described below in section 4.3.

		Utility at latency of		
Page number	**Threshold(i) (sec)**	**6 sec**	**10 secs**	**16 sec**
1	16	0.63	0.38	0
2	16	0.63	0.38	0
3	16	0.63	0.38	0
4	16	0.63	0.38	0
5	10	0.58	0.30	-0.12
6	10	0.55	0.25	-0.20
7	10	0.53	0.21	-0.26
8	10	0.51	0.19	-0.30
9	10	0.50	0.17	-0.33
10	10	0.49	0.15	-0.36
11	10	0.48	0.14	-0.38
12	10	0.48	0.12	-0.40
13	10	0.47	0.12	-0.42
14	10	0.47	0.11	-0.43
15	10	0.46	0.10	-0.44
16	10	0.46	0.09	-0.45
17	10	0.45	0.09	-0.46
18	6	0.43	0.05	-0.53
19	6	0.41	0.01	-0.59
20	6	0.39	-0.03	-0.64

21	6	0.37	-0.06	-0.69
22	6	0.35	-0.08	-0.73

Table 3: Utility values for different latencies over time.

4.3 Task expectations and latency requirements

Although we did not specifically set out to measure the difference in latency based upon task differences, we nevertheless found significant differences in tolerance and this was confirmed during focus group discussions. The importance of the finding is that users' expectations are different based on their beliefs about the complexity of the task, the slowness of access of remote information, or the effects of caching. We can apply this to web site performance by classifying requests by the kind of access type and then apply appropriate deadlines to the task. To establish the expectations of users, the task sets for each site can be profiled through user testing. A table can be established that maps quality ratings to the latencies of specific tasks. This information can be used in conjunction with duration information as well.

The association of tasks to deadlines can be encoded in the URL. There are several ways that this can be done. The simplest is to embed this information in parameters for the URL so that servers can classify requests with minimum performance penalty. For example is a URL that our participants believed could take some time to process because it retrieves a shopping basket. On the other hand, the main home page http://www.shopping.village.hp.com was one they felt should load up immediately. We can associate a deadline with the URL by to passing this information as a parameter that can be quickly parsed and does not require consulting a table of URLs for the appropriate deadline. For deadlines that can be calculated once such as task based deadlines the URLs can be included in the parameter, for example:

http://www.e-shopping.com/lots_of_goodies?task_deadline=3

The parameter could also specify a *task-type* flag that can be interpreted by the servers, for example:

http://www.e-shopping.com/lots_of_goodies?task_type=fast
http:/www.e-shopping.com/calc_cart_costs?task_type=calc

This *task-type* flag technique allows the deadline to calculated by the task using other information, such as client class of service or duration of the session for this client. For sites that do not maintain client based session information, the URLs can be dynamically generated to include session duration as well as task deadline parameters, for example by incrementing counters embedded in the URLs. Many e-commerce sites dynamically generate all URLs in each page and include client specific parameters, so there may already exist a framework under which these parameters can be conveniently added.

4.4 When to send feedback?

Using our measures of the latency users will tolerate, we can modify servers to provide feedback to users. Our results have shown that providing information concerning the processing of a request can significantly increase users' tolerance of poor QoS. The experiments in incremental loading show that people are willing to classify the service as high for up to 39 seconds instead of 5 if they have some notion that progress is being made. If the request has not completed within 5-10 seconds the user should be sent some indication that the request is still being processed. At this threshold the user is unsure whether the request has not been received correctly, if the site is down, or if the transaction has failed. The feedback to the user can be delivered as a multi-part HTML reply and can keep the client informed of the progress. It can even include boiler plate information that will be included in the final complete response. Without knowledge about when feedback is or is not to be expected, many users abandon long running requests. Informing the user if their request will be delayed above the established threshold for tolerance implicates that the QoS of the task as a whole is seen as better:

`I think it's great...saying we are unusually busy, there may be some delays, you might want to visit later. You've told me now. If I decide to go ahead, that's my choice'

4.5 QoS Auditing

While a site may make every effort to provide high quality of service, the measurement and interpretation of actual performance is essential. Using the thresholds of acceptability of response time, the objective performance to Web server monitoring data such as log files can be interpreted. For example, if the response times are less than 5 seconds 95% of the time and we know from our study that this constitutes high performance, then we can conclude that users experience high QoS 95% of the time at that site. Web servers can categorize and report very specifically for how many users it delivered high, medium, and low qualities of service, over time based on transaction volume. Servers can even identify specific pages for which service has deteriorated. For example, if the QoS was high during all the page viewings up to checkout, but checkout drops to medium or low, this change in performance can be noted. This can point out aspects of Web server and application servers that may need additional capacity to boost performance. The session QoS can also be reported and correlated to specific tasks that experience have poor performance at the site. The QoS auditing information could also be used to give feedback to the users. For example, if the server maintains historical data about site access patterns, then when it sees itself under a heavy load it can compare that load to typical load measures for the particular time of day/week/month. It can then provide feedback to the user about what level of service they can expect to get. The server might even be able to suggest a time when the server is less likely to be so heavily loaded, or predict what the response time will be. This interpretation of log and real time monitoring data is critically important as it provides a user's perspective of performance. The interpretation of objective thresholds of performance can also be used to decide when a site should be upgraded if the goal of the enterprise is to assure high QoS. Without these objective measures, the enterprise is only presented with absolute server response numbers, with no way to associate this data with user perceptions of QoS.

5. Conclusions and future work

This study was designed to investigate users' requirements for Internet QoS. We have shown that:

- The task in which users are engaged, the length of time they have been interacting with a site, and the method of page loading affects the acceptability of QoS.
- Tolerance of delay is influenced by the conceptual models users have of how the system works.
- Poor web-site performance leads to poor corporate image and often compromises users' conceptions of the security of the site.
- The findings of users' perception can be interated into server design and therefore result in QoS controls that reflect users' perception of quality.

We have shown that users' behavior in reaction to the level of QoS can be objectively quantified. Our findings have outlined a set of objective thresholds that reflect users' subjective assessments of quality. We were also able to identify salient parameters to a utility function. This function can be used to predict users' dynamic reactions to the QoS they receive. Predicting such reactions is a crucial step in accommodating user demand.

Our study focused on a Web-shopping task and the implications for server performance. We now have data to modify the server to give performance based on the absolute measures of latency for high, average, and low quality of service. To further validate these results, empirical work is needed to test these technical assumptions, for example observing user interaction with a modified server, to determine if a server that consistently meets the objective thresholds for high QoS is in fact perceived by users as providing a high QoS experience.

Our experience with duration of Web site interaction indicates that thresholds of acceptability change over time. A more precise mapping of these changes is needed. Again, it would be interesting to modify a server to improve the completion times of pages as the session time increases. Ideally, we would then be able to analyze the logs of the site and note that there was an increase in completed sessions and therefore successful buying operations.

There are a number of areas in which the study could be made more comprehensive. An obvious improvement would be a larger study with that includes female subjects. Female e-commerce shopping is growing rapidly and they may have different user perceptions which should be measured and incorpotated into the findings. Our study covered one specific e-commerce site, but certainly the experiment needs to be repeated with a variety of sites to be confident of the generality of the results. This study was also specific to a Web shopping task. Further study of users' perceptions of QoS should investigate the validity of our findings in different genres of Web usage, such as entertainment. The combination of results from different genres would make it possible to create more comprehensive conceptual models to predict how tolerance changes over the length of a session. Our research represents an important first step in identifying that such a relationship exists, and therefore indicating the need for technology to meet this need.

Acknowledgements

We thank Ilja Bedner for invaluable assistance with system configuration, Sharad Singhal, Ed Perry and Ilja Bedner for their help with experimental design, and members of HP Labs for participating in pilot studies. We also thank the WWW9 reviewers for their helpful comments.

References

1. Abdelzaher, T., and Bhatti, N. "Adaptive Content Delivery for Web Server QoS". Proceedings of IWQoS`99 (London, May 1999).
2. Alameida, J. et al, "Providing Different Levels of Service in Web Hosting Hosting", Proceedings of the Internet Server Performance Workshop, March 1998.
3. Banga, G. and Druschel, P., "Resource Containers: A new Facility for Web Content Hosting", Proceedings of the Internet Server Performance Workshop, March 1998.
4. Bernet, Y. "A Framework for End-to-End QoS Combining RSVP/Intserv and Differentiated Services". IETF (March 1998).
5. Berst, J., "Bandwidth progress report". Available at http://www.zdnet.com/anchordesk/story/story_1384.html
6. Bhatti, N., and Friedrich, R. "Web Server Support for Tiered Services". IEEE Network (September/October 1999).
7. Bouch, A., and Sasse, M.A,. "Network QoS: What do users need" IDC`99 (Madrid, September 1999).
8. Bouch, A., and Sasse, M.A. "It ain't what you charge it's the way that you do it: A user perspective of network QoS and pricing.", Proceedings of IM'99 (Boston MA, May 1999).
9. Bouch A, et al., "Quality is in the Eye of the Beholder: Meeting Users's Requirements for Internet Quality of Service". Proceedings of ACM conference on Human Factors in Computing Systems (CHI 2000), to appear April 2000, the Hague, The Netherlands.
10. Boyer, D.L., Pollack, J.G., and Eggemeier, T.F., "Effects of Aging on Subjective Workload and Performance: Determinants of Age Differences in Cognitive Performance.", Proceedings of Human Factors Society 36th Annual Meeting 1, (1992).
11. Card, S.K., Moran, T.P., and Newell, A., The Psychology of Human-computer Interaction, Lawrence Erlbaum Associates, Hillsdale, NJ, 1983
12. Cullinane, P., "Ready, set, crash". Telephony, (Nov 1998).
13. Dunlop, M.D., and Johnson, C., "Subjectivity and notions of time and value in interactive information retrieval.", Interacting with Computers 10, 1, (1998).
14. Fishburn, P.C., and Odlyzko, O.M., "Dynamic behavior of differential pricing and Quality of Service options for the Internet", Proceedings of ICE-98, ACM Press.
15. Fox, R. "News Track", Communications of the ACM, (May 1999), 9-10.
16. A., Stahl, D.O., and Whinston, A.B., "Pricing of services on the Internet", http://cism.bus.utexas.edu/alok/pricing.html.
17. Hogan, M., "The first ever report on the top 100 e-commerce businesses and the secrets of their success", PC Computing Magazine, June 8, 1999.
18. Jakob Nielsen's, "Top Ten Mistakes of Web Design" http:/www.useit.com/alertbox/9605.html
19. Kawalek, J., "A user perspective for QoS management", Proc. QoS workshop, 16th September 1995, Crete, Greece.
20. Keller, J.J., "Ex-MFs managers plan to build global network based on Internet", Wall Street Journal January 20, 1998.
21. Kokotopoulos, A., "Subjective assessment of multimedia systems for distance learning", Proceedings of the. European Conference on Multimedia Applications, Services and Techniques, Milan, May 1997.
22. Mackie-Mason, J.K., and Varian, H., "Economic FAQs about the Internet". In McKnight, L.W., and Bailey, J.P. (eds.). Internet Economics. MIT Press, 1997.
23. Morgan, K., Morris, R.L., Macleod, H., and Gibbs, S. "Gender Differences and Cognitive Style", Human-Computer Interaction. Proceedings of EWHCI`92, 1992.
24. Myers, B. A., "The Importance of Percent-Done Progress Indicators for Computer-Human Interfaces", Proceedings of CHI"85 (San Francisco CA, April 1985)
25. Newman, W., and Lamming, H. Interactive Systems Design. Prentice Hall. 1995.
26. Nielson, J., Usability Engineering, AP Professional Press, Boston MA, 1994.
27. Perry, C, "Travelers on the Internet, A survey of Internet users", Online 19, 2, (1995).
28. Ramsay, J., Barbesi, A., and Preece, J., "Psychological Investigation of Long Retrieval Times on the World Wide Web.", Interacting with Computers 10, 1, (1998).
29. Rogers, Y., "Evaluating the Meaningfulness of Icon Sets to Represent Command Operations", Display Based Systems: Procs. HCI'86 Conference on People and Computers II, 1986 p.586-603
30. Wang, Z., "USD: Scalable bandwidth for differentiated services", http://www.ietf.org/drafts-wang-00.txt
31. Watson, A. Sasse, M.A., "Multimedia conferencing via multicast: Determining the Quality of Service required by the end-user". Proc. International Workshop on Audio-Visual Services over Packet Networks (AVSPN) 15-16 September 1997, Aberdeen.
32. "ETE Watch for Web Browsers. http:/www.candle.com/etewatch.
33. "Graphic, Visualization, & Usability Center's WWW User Surveys", http://www.cc.gatech.edu/gvu/user_surveys
34. HP Shopping Village. http://www.shopping.hp.com.
35. "The Internet, Technology 1999, Analysis and Forecast", IEEE Spectrum (January 1999).
36. "Network Reliability Steering Committee Annual Report 1998". http://www.nric.org

1. Current address: Department of Computer Science, University College London, Gower Street, London, WC1E6BT, A.Bouch@cs.ucl.ac.uka
2. To illustrate key points, we have included quotes taken from the verbal protocols and focus groups.