Understanding How Users View Application Performance

Net Forecasts – Peter J. Sevcik
BCR Volume 32, Number 7
July 2002

Everyone wants to attract users to their Web site and encourage them to return. This is true for *any* application regardless of the commerce implications. Much has been written about what constitutes the "user experience," and most of the emphasis has been on page-response time – how long it takes for a page to load onto the user's PC.

But research on this issue is all over the map; desired response times vary dramatically. As a result, it's been tough to answer fundamental questions: How fast is fast enough? How slow is too slow?

So, my colleagues and I tried to tackle this problem. We developed a model that appears to resolve the conflicting data on how users feel about their interaction with computers and the Web.

In our model, each user interaction constitutes a "task," and we found three zones of duration that represent how users feel about each task: satisfied, tolerant or frustrated. Users set their satisfaction threshold differently depending on their expectation of the amount of time they're willing to invest in the outcome of the task.

Although the research studies cited below refer to Web pages, these findings apply to all applications, and it's important to emphasize that our interest in this issue is not academic. Enterprises are spending considerable time and money trying to develop proper performance policies.

Zone of Satisfaction

Let's say that the value – i.e., the time – for a "satisfied" performance zone is under *A* seconds. According to a number of studies, a user is "satisfied" when he/she is not conscious of the time it is taking to load the page. Determining the value of *A* has proved elusive, despite its importance.

Back in 1968, research by IBM's Robert Miller showed that users remain completely engaged in computer dialog, as long as the response time is less than 1 second [1]. That led to the 1-second response-time target for most IBM mainframe applications. Miller also found that after 10 seconds, the user's mind starts to wander and he/she loses productivity with respect to the computer tasks.

A more recent study by Nina Bhatti, *et al*, looked at how users reacted when they were directed to use the Web to configure and buy a PC [2]. That research yielded a similar result. There was a definitive negative shift of perception at 10 seconds. Interestingly, the maximum tolerance for the page-download time decreased from 10 to 4 seconds as the session progressed toward the goal of buying the computer.

Those findings are more or less compatible with Zona Research's often-quoted "8-second rule" [3], and also with last year's Gomez survey that found that users preferred to have a page load in less than 10 to 12 seconds, depending on the type of site they are accessing. But just when we were beginning to sense a consensus regarding the value of A, we stumbled across counter examples.

For example, Jarad Spool of User Interface Engineering, found that Web users' perceived speed of a site did *not* correlate to the actual download time [4]. Instead, perceived speed correlated strongly with successful task completion.

Spool studied a number of users who visited 10 sites, and chose the functions that interested them the most. At the end of the test, each user rated the speed performance of each site. One site, about.com, was deemed the slowest by users, when it was in fact the fastest (8 seconds); meanwhile amazon.com was perceived as being the fastest site, even though it was actually the slowest (36 seconds).

Why this apparent contradiction? We believe it comes down to defining "elements" -- the few words or numbers that a user is looking for or will spend time reading – and we believe that a user expects to spend about 4 seconds per element. About.com is full of short phrases that the user must click on to actually get to the next level, so a

user is forced into a 1-element interaction, which the user "expects" to take 4 seconds, not eight. By contrast, Amazon.com engages the reader with many elements of interest, and so the user spends extra time reading each page; they're content with the experience even though 36 seconds have passed.

This also explains why the PC buyers in the Bhatti study were satisfied with 10-second interactions in the beginning of the process, when there was a lot of information to read (finding and choosing options). But as the transaction progressed to the point where each page represented a single element of information (the price, order confirmation, etc.), they wanted shorter, 4-second downloads

We believe that users pre-set a "budget" of time for a page, based upon the expected number of elements they will read. That budget is 4 seconds per element for a Web interaction, and the budget might be set before the user ever clicks on the page or it may form as the page loads.

By selecting an average from evidence cited above, the zone of satisfaction in North America must be 10 seconds or less for a typical page with 2-3 elements.

Zone of Tolerance

The next level down begins when the page-load time exceeds that experienced in the zone of satisfaction. Again, studies show that this is when the user becomes aware of the fact that the page is taking time to load; time becomes a factor in a user's satisfaction.

And as we all know, what starts as simple awareness of the passage of time, slowly builds into annoyance. We don't know much about a user's mental state or behavior when he/she is in this zone, but we do know that many users tolerate this page-load duration without quitting. We also know that there is the wide band of time between when a user is no longer satisfied and when he/she becomes downright frustrated. When that happens, the user moves into the next zone.

Zone of Frustration

The third zone is where things get ugly; the user has waited to the point where he/she is significantly frustrated. Let's call the threshold into this zone *B* seconds.

Judith Ramsay, *et al*, looked for the value of *B* in a fascinating study, which showed that users' level of interest in content radically changes at 41 seconds or longer [5]. Nina Bhatti's users were slightly more patient – they indicated frustration at 39 seconds [1], and another researcher, Paula Selvidge, ran experiments that showed users began experiencing significant frustration when pages loaded in 30 seconds or longer [6].

Our experience with mission-critical enterprise applications is that the original IBM 1-second per element still applies, and that most production tasks are single-element functions. We have seen people doing production data entry completely lose their work cadence – i.e., productivity is severely impaired -- if tasks respond in more than 4 seconds.

So, we conclude that *B* is consistently 4-times *A*. This means that a North American Web user accessing a North American Web site generally will become frustrated if the typical page loads in more than 40 seconds.

Of course, there are times when users will persevere, but any company using the Web to conduct commerce or court customers has to engineer their sites to avoid the 40-second threshold. Actually, when you think about it, 40 seconds is longer than we tolerate in almost any interactions with technology. If your car took 40 seconds to start, wouldn't you think something was wrong? If your bank's ATM machine needed 40 seconds to verify your account or complete a single transaction (deposit, withdraw), would you think the bank was doing an OK job?

Conclusion

Obviously, there is no single performance rule that covers all situations. But we believe that 10 seconds is the upper limit for user satisfaction, and 40 seconds is the limit at which performance becomes intolerable. These rules of thumb can be used as a starting point when evaluating Web application performance.

This method applies equally to all vertical markets. The key is the number of useful elements on the

Web page, and some industries, by their very nature, have little to say. For example, financial industry websites typically don't provide a lot of text; users click on a page to get one or a few numbers – a stock quote, account balance, etc. -- so all of their zones are shorter.

The threshold values vary depending the nature of the task, which also governs the number of elements the user is likely to want to read. The thresholds also change as a user changes what he/she is doing. However, the thresholds do *not* change based on the type of site being visited. This means that thresholds vary within each Web site, an important, and heretofore, unnoticed finding.

Contrary to conventional wisdom, variations among users are not as important in determining performance-tolerance levels, as are variations among the tasks that users undertake. The mental processes by which we interact with machines are grounded in the way we work, and we humans aren't evolving at the rate of Moore's Law.

I welcome your response and comments on this model of behavior; I can be reached at peter@netforecast.com. I also invite anyone interested in this topic and the performance of applications on a network to attend my tutorial on the subject at NGN on October 14, 2002 in Boston.

References

[1] "Response Time in Man-Computer Conversational Transactions", R. B. Miller, *Proceedings of the AFIPS Fall Joint Computer Conference*, 1968.

[2] "Integrating User-Perceived Quality into Web Server Design" by Nina Bhatti, Anna Bouch, Allan Kuchinsky, *9th International World Wide Web Conference*, May 2000.

[3] "The Need for Speed" by Zona Research, July 1999.

[4] Jared M. Spool, An interview with Jared Spool of User Interface Engineering, conducted by John Rhodes for *WebWord*, July 2001.

[5] "A psychological investigation of long retrieval times on the World Wide Web", Judith Ramsay, Alessandro Barbasi, Jenny Preece, *Interacting With Computers*, Elsevier, March 1998.

[6] "How Long is Too Long to Wait for a Website to Load?", Paula Selvidge, *Usability News*, Wichita State University, July 1999.

Peter Sevcik is president of NetForecast in Andover, MA, and is a leading authority on Internet traffic, performance and technology. He has contributed to the design of more than 100 networks, and led the project that divided the Arpanet into multiple networks in 1984, which was the beginning of today's Internet. He can be reached at peter@netforecast.com.