How Fast is Fast Enough?

Net Forecasts – Peter J. Sevcik
BCR Volume 33, Number 3
March 2003

A few months ago, I described a foundation for thinking about user perceptions of Web performance in three zones: satisfied, tolerating and frustrated (see *BCR*, July 2002, pp. 8-9). I noted evidence that suggests the upper limit of the "satisfied" zone for browsing the Web is 10 seconds.

That prompted strong criticism from Peter Christy who, in a subsequent *BCR* article, argued: "We have to stop thinking of 10-second response as being OK in any remote sense of the word, and relearn the performance understanding of 20 years ago. The right goal is a fraction of a second." (See November 2002, pp. 46-47.)

I have never contended that there is one – and only one – number that is the "right" performance target. In fact, I noted that for "mission-critical enterprise applications…the original IBM 1-second per element still applies."

But to help clarify the situation, I want to describe the methodology we use to set application performance targets. There are two key factors: The number of elements viewed and the repetitiveness of the task.

The number of elements viewed is the number of items, fields, paragraphs, etc., that the user is *expecting* to study or look at just *ahead* of their arrival. The user sets wait-time expectations based on the amount of time he/she plans to spend with the content.

The user also interacts with applications, at a certain pace driven by how repetitive each task is within the overall session. Some are highly repetitive, others require the user to think and make choices before proceeding to the next screen. There's a subjective continuum, from simple data entry to casual browsing; after all, they call it a *browser* for a reason.

Table 1 shows a matrix of response-time limits to the satisfaction zone, based on the intersection of these two metrics. Note that the user can be satisfied within a range of 1 to 16 seconds; the "proper" response-time target is based on what the user is doing and is not a single number for all circumstances. While there is a large body of evidence that supports these conjectures, there is no formal proof of this exact approach, so feel free to change the specific time values in the table based upon your understanding your users' needs. And very unique situations, like twitch games, have a different matrix.

Table 1 – Alternative Limits of the User Satisfaction Zone (sec)

	Number of Screen Elements User Views			
	1	2	3	4
Task Repetitiveness				
Very Low	4	8	12	16
Low	3	6	9	12
High	2	4	6	8
Very High	1	2	3	4

10 Seconds vs. 1 Second

When first presenting this methodology, I stated that a user typically looks at about 2.5 things (sometimes 2, sometimes 3) in a Web page at a slow pace. Therefore, I selected the value midpoint between columns 2 and 3 in the top row of Table 1–10 seconds. So, the typical Web page should load in less than 10 seconds.

Peter Christy argues that interactions with applications – even across the Internet – should operate at less than 1 second. His article would lead us to believe that there is something wrong with any interactions that are slower than 1 second. Lets compare the implications of these opposing views.

In my January 2003 column, I published the results of a study that measured typical business site pages as seen by broadband users across North America. Two different measurement services supplied data, but produced very similar results. About 2 percent of the users saw pages arrive in less than 1 second, while 82 percent had them arrive in less than 10 seconds. So, the fact of the matter is that very few broadband and none of the dial-up users currently experience the fast performance Peter Christy proposes.

However, we also know that, last year, more than 200 million people (Nielsen//NetRatings) accessed many of the 2 trillion Web pages indexed by Google. Furthermore, just in the U.S. alone, 67 million households purchased $48 billion of goods and services over the Internet in 2002 (Jupiter Research). The business-to-business market is estimated to be an order of magnitude larger than the retail market.

It is impossible to accept that all this activity and commerce was generated by a user population where only 2 percent of the users were satisfied with performance. Either performance does not matter, or the 10-second target, which encompasses 82 percent of the population, must be the operative reality.

Setting Performance Targets

I am not saying that all is well with the world; many Web users see terrible performance. About 1 in 10 broadband users in North America see a typical Web page load in 27 to 42 seconds depending upon the measurement service. So, a significant number of users are frustrated with performance each day.

Furthermore, some users and applications need response times faster than 10 seconds. This is particularly true for enterprise applications, whether they are operating over the Internet or over private networks.

The key first step in setting performance targets is for enterprises to investigate how their applications are used: How do real users perform work in real-life network situations? Here Peter Christy and I agree: Enterprises must measure performance by multiple means and over time to truly understand the situation. There is no substitute for hard data.

At some point, application performance will need improvement, and there are a host of strategies, technologies and services that can help in that effort. But, doing so requires that management make decisions that will cost money and disrupt some of the infrastructure in to achieve the new performance goal.

How will you know when you are done? When will you stop spending more money to improve performance? You need targets that are based on a rational, unified understanding of your user population. The methodology of Table 1 is one such approach. The targets not only must define the boundary of satisfaction but must also define the boundary of "tolerance." Finally, the targets must be defined for portions of the user population (e.g., 90 percent of our users will see less than 6-second response time).

Without specific targets, the enterprise will never finish a "make it faster" project. To adopt Peter Christy's simple goal – "the right answer is a small fraction of a second" – would condemn enterprises to non-stop spending on the problem.

The Internet is Transformational

Christy is correct, however, when he says that the Internet gives access to corporate information in a way that changes how businesses operate. Managers, customers and partners can and, often, *must* directly access enterprise information to be a part of a business process. Information intermediaries like secretaries and business agents are being eliminated.

Where Peter Christy and I disagree is on how this impacts response-time targets. The conventional wisdom shown in his SAP example is that the clerk used the application more than the finance manager. Therefore the application could be permitted to run slowly because if the clerk, "found it slow and cumbersome, tough luck, find a new job if you want." But in today's real world, managers use SAP, so Christy says it had better run faster commensurate with the population shift.

The opposite is true. For example, when production workers enter data into a business database, they usually operate in the lower left-hand corner of Table 1. The job is tedious and repetitive, but it's also demanding, because the clerk must be fast and accurate.

By contrast, when a manager asks for a report of the processed data, he/she will spend a lot of time with the result, so they are in the upper right corner of Table 1. The manager clicks for a summary report to study and interpret the report in order to make a decision – that's what makes him/her a manager.

The Internet is transforming this world of clerks and managers. The clerk's job is being replaced by software, like supply-chain management, logistics integration, order fulfillment, etc. Today, when you order a product online, you're increasingly likely to also get a shipping number and URL so you can perform your own tracking. No human entered that tracking number, and steps that used to be performed by several clerks were eliminated. I agree with Christy that the ratio of on-line clerks and managers is moving towards more managers, but I believe that means the response-time targets for those clerks that remain may get lower (i.e., faster), while the targets for managers that are climbing in number become higher (i.e., slower).

Invitation to the Dialog

Christy and I both care about this topic and think that it's important for enterprises to get this right, and I'd like to broaden the participants in this discussion. So, NetForecast, with help from *BCR*, is launching a survey of performance from the enterprise perspective: What level of performance really matters?

You are invited to participate. Your information will be kept confidential. Just drop me an email (peter@netforecast.com) stating that you would like to be interviewed for the project. Participants will receive a report of the findings and a summary will appear in a future column. Please, join in!

Peter Sevcik is president of NetForecast in Andover, MA, and is a leading authority on Internet traffic, performance and technology. Peter has contributed to the design of more than 100 networks, including the Internet, and holds the patent on application response-time prediction. He can be reached at peter@netforecast.com.