

A Methodology for Determining Response Time Baselines: Defining the “8 Second” Rule

By Charles Hoover
Operations Manager
CARFAX, Inc.

Abstract

For a long time the 8-second rule has been the norm for setting response time on web pages. But how accurate is this rule in our new high-speed, broadband era of the Internet? By looking at the previous research done on user expectations and collecting response time data from a variety of data sources, it has been possible for us to come up with basic baselines. Then utilizing the Application Performance Index (Apdex) we were able to compare the response times of various pages to see how well they performed.

Introduction

A couple of questions that we at CARFAX have always been wrestling with ever since we went on the web have been “what is a good response time from our web site?” and “what is a bad response time from our web site?” We have followed the “8 second rule” from the beginning, which says that any response time greater than 8 seconds will cause us to lose customers, but no one seemed to have an answer to what was a “good” response time. The 8-second rule was the industry standard, but as we purchased more equipment, more bandwidth and more human resources, in order to make sure we don’t violate this rule, we’ve always wondered if it was truly an accurate baseline to measure against or if there was a better way to measure and monitor response time.

After some research into user perceptions and understanding our customers better, we think we have come up with a methodology that enables us, in general, to set good baselines for response time and thresholds for poor response time.

The “8 second rule”

The “8 second rule” is the accepted rule-of-thumb for web site response time ever since the World Wide Web began. It has been just accepted as fact that after 8 seconds a user will stop waiting for a web page to load and move on to another site, probably a competitor if you are in e-business. This standard goes back to the 1968 Robert J. Miller paper for IBM, “Response Time in Man-Computer Conversational Transactions” [Miller] in which he describes the three thresholds for user attention:

- 0.1 Seconds – user sees this as instantaneous
- 1.0 Second – uninterrupted user actions, they will notice a delay but can continue to work
- 10 Seconds – limit for keeping user’s attention

Miller identified a two-second response time as ideal. These standards were later used as business shifted from mainframes to the PC and later the web.

The “8-second rule” was adopted later after the paper, “Worth the Wait?” by Peter

Bickford reported that half the users abandoned a page after 8.5 seconds [Bickford]. There has been other research conducted by Jan L. Guynes that shows users actually become stressed after 8 seconds [Guynes]. But even this research cautions not to generalize their research to other computer related tasks.

So while there seems to be some evidence that 8 seconds is a good rule-of-thumb for response time when it comes to some computer tasks, it is by no means definitive. As Chris Loosely points out in his article, "When Is Your Web Site Fast Enough?", while Miller's paper identified the important behavioral thresholds that all humans share, and Bickford and Guynes adds some evidence to support it, each user's behavior will depend on prior experience either with the web site in question or with the web as a whole [Loosely].

In his article, "Do Interface Standards Stifle Design Creativity?" Jakob Nielsen states his *Law of Web User Experience*: "users spend most of their time on other sites" [Nielsen], implying that a typical user's expectation of response time on your web site is based on their experiences on other web sites. This is an interesting conjecture because it indicates that in order to meet the expectation of a user coming to your website is not based on how fast your web site has been, but how fast everyone else's web site has been. So monitoring your own response time is only meaningful in terms of monitoring other websites and comparing it to those other sites.

Beyond the 8-second Rule

If we use Nielsen's logic that a user's perception of appropriate response time for our web site is based on the user's experience at other web sites, then we should gather response time data from other web sites and compare our response

times to them. This is fairly easily to accomplish thanks to www.keynote.com, which keeps response time data, free to the public, on a weekly basis for the following companies:

Amazon	Costco	JCPenney
Office Max	Target	BestBuy
Eddie Bauer	Sears	Office Depot
Walmart		

Keynote then ranks the response times and publishes the top three for the week, on its web site, along with the average for the week for all the sites. The response timings are the total time it takes a user to: log in, search for an item, add it to a cart and check out.

Note: Keynote has made several changes to its web site since this paper was first written and the data collected, but this data should still be available at:

http://www.keynote.com/solutions/mm_public_services.html

Unfortunately, Keynote has temporarily discontinued the data collection that was used for this paper. They do have other data for Retail sites, based on dial-up speeds, and home page response times for Government and Business sites. Another source of free data for web response time can be found at www.alexa.com.

Analyzing this data from August 2005 to January 2006 we find that the average for these web sites is 10.45 seconds, with the average for the top three being 7.42. Assuming 4-5 steps for each transaction (home page, search page, product view, order page and credit card validation), we see that the top three sites fall well within the Miller's ideal two-second per interaction, ideal response time. All the sites are close to the two-second ideal, averaging between 2.6 to 2.09 seconds per transaction step depending on a 4 or 5 step transaction, and

1.86 to 1.48 seconds for the top three. Thus following Nielsen's Law of Web User's Experience, we can assume that users are expecting a response time of around two-seconds or faster.

Meeting Expectations

Our original question is still valid, with an added caveat: Should our response time be close to others web response times and is this best way to meet user expectations concerning our response time? These companies have a lot of resources to use to make sure that their web sites provide excellent response times, while we may not be so fortunate. So is there a better way to meet our customer expectations without spending ourselves into a hole?

First some explanations of what CARFAX does. We are a Vehicle History Service, which means that we keep historical data on cars. By going to our web site, and entering the Vehicle Identification Number (VIN) a customer can produce a detailed history for any car back to 1981 when the VIN was first used.

The next thing to keep in mind is the steps for a transaction. For most users a typical transaction breaks down as follows:

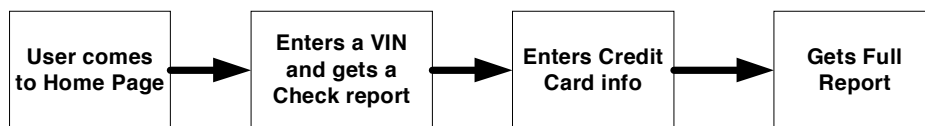


Figure 1

The user comes to the Home Page, and then will enter a VIN into our Check Report. The Check Report is a free version of the Full Report but which only indicates what information will be available on the Full Report. For most users this may be as far as they will go because the Check Report will tell them if there are any major

problems with the vehicle that they are interested in. If they decide they would like the Full Report, then they enter their credit card information. Once it is processed, then they get the Full Report in their browser, at which point they can have it e-mailed to them or they can save it as a PDF file.

Because of the way we've designed our processes, the response time for a Check Report is roughly the same time as for a Full Report. This is because when any user enters a VIN, the information to produce a Full Report is returned; the only difference is what gets displayed on the page. Therefore, the response time for a Check Report is very close to the response time for a Full report, so when we collected our initial data, we only captured response time information for our Home Page and Check Report.

The Home Page

For new users, the response time for our Home Page is very important. This is where Nielsen's Law of Web User's Experience really comes into play. Because they have nothing to base a response time experience to our web site on, they will expect what they have experienced at other sites. Therefore we can use the Keynote results of an average of 2 seconds as a baseline to measure our Home Page against.

We captured external response times utilizing the service provided by Quantiva. Unfortunately, Quantiva ceased to exist on September 2005 so we only have data for the previous year from January to October (the date they finally turned their servers off). Quantiva polled our web site every 15 minutes, simulating a new user to avoid any

caching issues. They also simulated a broadband user. Their collectors were set up in Newark, New Jersey; Chicago, Illinois; Dallas, Texas and San Jose, California and utilized AT&T, Uunet, Internap, and Saavis as Service Providers. The data was collected daily and stored as separate data files. For this analysis, these daily files were averaged daily and then merged together to form a single data file.

From Chart 1 we can see that, aside from a few spikes, we were well within the 2-second, optimal times for most of the year. The longer periods of higher response time indicate weeks when we were running marketing tests. These marketing tests were designed to see if changes to the Home Page generated more traffic. It

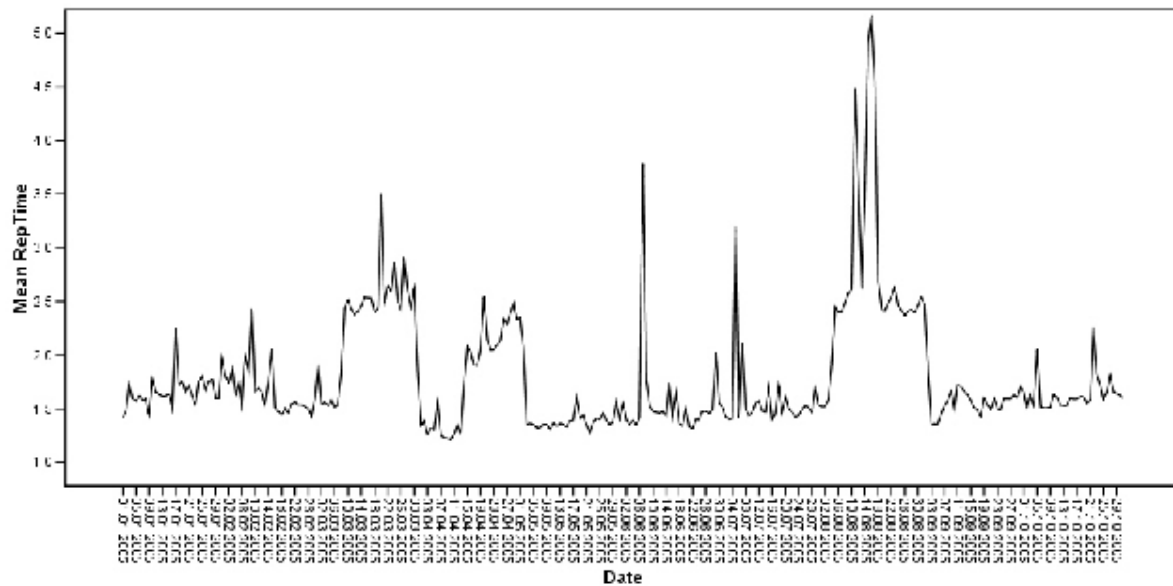


Chart 1: Daily Response Time for Home Page

Histogram

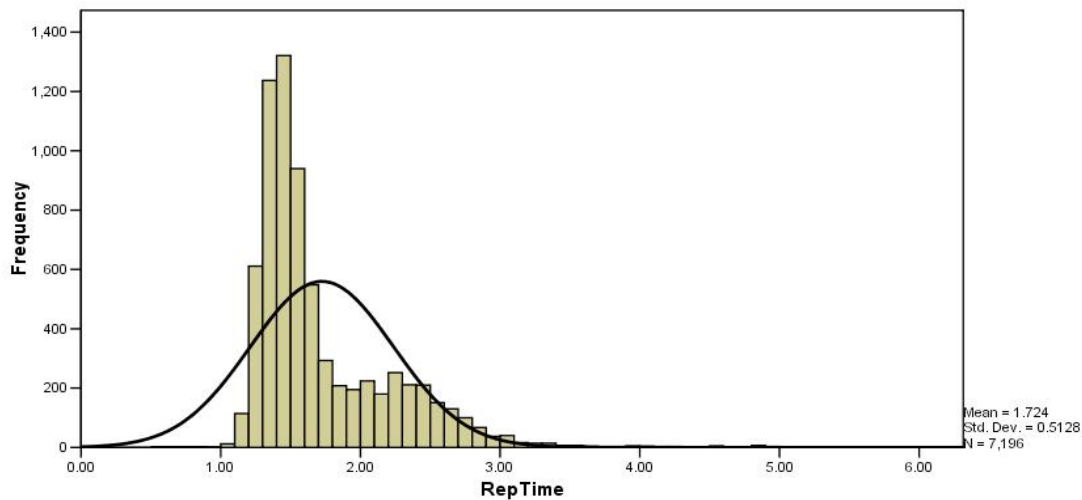


Chart 2: Normal Distribution of Home Page Response Time

should be noted that it was these marketing tests that were the catalyst for asking about what are suitable response times. Still, even these longer spikes often didn't go above 3 seconds, which should have been an acceptable response time.

Looking a little deeper we can see that, on average, the response times for the Home Page are actually skewed towards a one-second response time rather than towards two-second or higher response time.

The Normal curves shown in Chart 2 demonstrate that our response time is skewed toward 1 or 1.5 seconds, beyond what would be expected from just random frequency. For this chart, the outliers for response time were deleted to get a better sense of the shape of the frequency curve.

If we follow Nielsen's logic that consumers will expect to get the response time that they get at other sites and Miller's logic that a two-second response time is optimal, then we can be fairly confident that our response time for the most part is meeting expectations.

The Check Report

Our average Home Page response time falls well within the accepted criteria set forth from the research done on user expectations, but what about the VHR itself? One option to answer this question would be to go back to the Keynote data and see how long other businesses on the web are taking to process transactions. Unfortunately, our business is different than the ones in the Keynote sample. These businesses are primarily brick-n-mortar stores with a web site, while we are selling information. Thus, our customer expectations for the final product they are getting will be a little different. For these brick-n-mortar stores, an argument could be made that their total transaction time should

include the delivery time of the product to the customer, since ordering is just the first part; the product needs to be shipped, then the shipping company needs to deliver the product before the customer can be fully satisfied. A closer model to our situation at CARFAX would be purchasing downloaded software since after the credit card validation goes through, the product can be downloaded and used immediately.

Also, there is usually a different expectation for response time when information is being queried from a database. In this case, an extremely fast response time may give the impression that there wasn't much data to query and therefore the report produced may be of questionable value. On the other hand, too slow a response time could indicate that our data isn't very well organized and thus may be prone to error. Research done by Bhatti, et. al. in their paper "Integrating User-Perceived Quality into Web Server Design", suggests that user expectations of response time can vary from 2-seconds to 39-seconds or longer depending on the right conditions. But they also indicate that there are many issues that affect user expectations, such as how often the user is on the Internet, if they are just browsing or specifically searching for something, etc. [Bhatti]. Utilizing the findings of Bhatti, et. al., for our purposes, and without more information to base a decision on, we can assume that the original "8-second rule" is probably the best indicator for what user expectations will be for returning information after a VIN has been entered.

Using the Quantiva data for Check Report response time, we have the following results:

Again, the data was averaged daily for the period from Jan 1, 2005 to October 30, 2005 and shows an average response time of 1.57 seconds. There were some higher spikes but nothing above 3.5 seconds

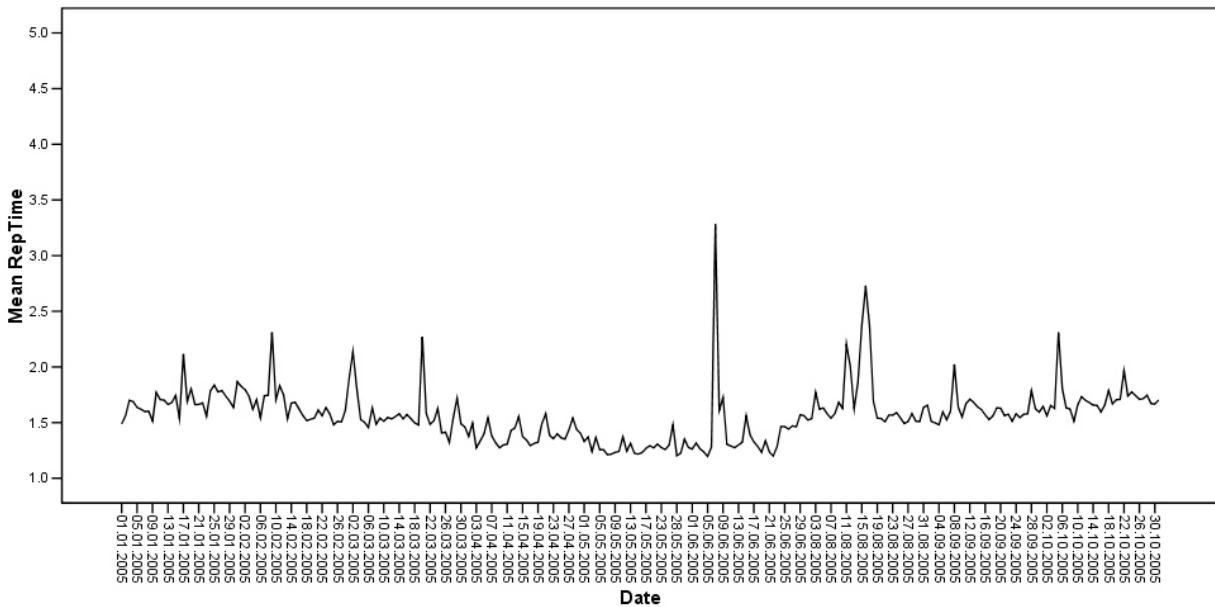


Chart 3: Daily Response Time for Check Report

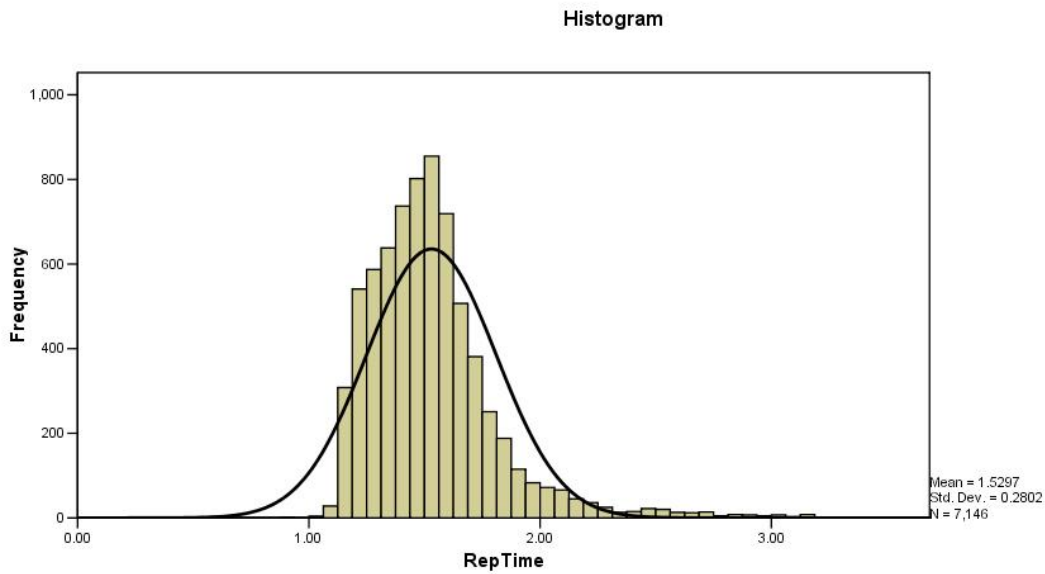


Chart 4: Normal Distribution of Check Report Response Time

which falls well within our expected 8-seconds. And when we look at the Normal Distribution in Chart 4, we see a tendency towards 1.5 seconds.

The Full Report may take slightly longer to display because of more information shown, but even if it doubles the time, it also is well within the 8 seconds that we are using as a rule of thumb for this web page's response time.

The Credit Card Validation

The only piece missing from the whole transaction equation is the Credit Card processing. This is another example of where prior experience will help set a customer's expectation. But getting "real" response times for credit card processing from other business will be tricky to say the least. If we utilize the Keynote data again we can postulate an average of 2.5 seconds for Credit Card processing since the entire four step transaction for the other sites was 10.45 seconds. Unfortunately, we did not collect regular response time data for Credit Card Validation from Quantiva, so we have nothing to analyze it against, but in the future we will know where to start.

Determining Poor Response Time

We've come up with ways to determine what is a "good" response time but is a "poor" response time? One way of doing this fairly easily is using the knowledge that went into producing the Application Performance Index or Apdex. Apdex is a methodology of utilizing what is currently know about customer satisfaction about response time and computing a single metric that has the same meaning for any web page, application or client-side tool. The index ranges from 0 to 1 with 1 being the best performance and 0 the worst. Apdex breaks the user perception of response time down into: Satisfied, Tolerating and Frustrated. Users then are either observed or self report on what is a satisfactory response time, and what is tolerable. These results are inputted into the Apdex formula:

$$\text{Apdex} = \frac{\text{Satisfied Cnt} + \frac{\text{Tolerating Cnt}}{2}}{\text{Total Samples}}$$

Where:

Satisfied Count = the number of responses that meet the satisfactory response time which is defined as: User maintains concentration, and performance is not a factor in the user's experience.

Tolerating Count = the number of response that meets the tolerable response time, which is defined as: User's concentration is impaired, performance is now a factor in the user experience or the user notices how long it is taking.

Total Samples = the total number of samples taken.

The Apdex Index can then be compared to their rating system which is as follows:

1.00 – 0.94	Excellent
0.93 – 0.85	Good
0.84 – 0.70	Fair
0.69 – 0.50	Poor
<0.49	Unacceptable

In determining what constituted Frustrated response times, the researchers who developed Apdex found that four times the tolerable response time was a very accurate measure of poor response time [Sevcik]. This also correlates with findings by others ([Bhatti] and [Selvidge]), who found that at around 30 seconds or 4X8, users abandoned whatever task they were working on. Therefore in our response time analysis, if we follow the expected response time for the Home Page of 2-seconds, based on the research and findings of Miller and Nielsen and the data from Keynote, the Frustrated point will be 8-seconds (2 X 4). For the Check Report we can assume an expected response time of 8-seconds since it is a database lookup and Bhatti showed with their research that users are willing to wait a little longer when data is being

queried. The Frustrated point will then be 32-seconds (8 X 4) or roughly again what we would expect from the research [Bhatti], [Selvidge], [Sevcik]. Finally the expected response time for the Credit Card Validation would be 2.5-seconds based on Nielsen and what most users expect by going to other web sites. The Frustrated point for this page would be roughly 10-seconds (2.5 X 4).

We can then monitor our response times for various web pages and compare them using Apdex to determine how well we are meeting Service Levels. If we assume our baselines are the Satisfied Level and then calculating our Frustrated Level by multiplying by four we can count our sampling and use the Apdex formula. Table 1 shows how we would set these zones:

Web Page	Satisfied	Tolerating	Frustrated
Home Page	0 – 2	2 to 8	> 8
Check Report	0 – 8	8 to 32	> 32
Credit Card Validation	0 – 2.5	2.5 to 10	>10

Based on these ranges, we can take our response time data and calculate our Apdex indexes for our Home Page and Check Report and determine what Apdex Ratings we can assign:

Home Page = 0.875 (Good)
 Check Report = 0.999 (Excellent)

Conclusions

By looking at the research already done on user perceptions of Internet response time and how their own expectations factor in, we can determine fairly accurately what baseline to use for a “good” response time. Some of this research is just common sense, such as Nielsen’s *Law of Web User’s Experience*, where our expectations of response time from a new web site are based on what we have experienced at

other web sites. It also comes from our own human nature where our anything longer than 10 seconds and we tend to lose focus on the task.

By then utilizing available data on response times for other businesses from such sources as Keynote, we can pretty accurately set our own baselines for what our expected response times are.

Then by applying Apdex we can measure various parts of our basic business transactions and compare the results to see where we need to spend our resources.

In the future, we will refine our efforts by focusing our research on the response times for other web sites within our industry and by expanding this methodology to other parts of our site. For this area of research as a whole, there should be more study of user expectations of database queries and products that involve pure information.

Bibliography

[Miller], Miller, R. B. “Response Time In Man-Computer Conversational Transactions”, Proc. AFIPS Fall Joint Computer Conference Vol. 33, (1968), 267-277.

[Bickford], Peter Bickford, “Worth The Wait?”, Netscape/View Source Magazine, (2000).

[Guynes], “Impact of Systems Response Time on State Anxiety”, Communications of the ACM, March 1998, Volume 31, Number 3.

[Loosely], Chris Loosely, “When Is Your Web Site Fast Enough?”, E-Commerce Times, (Oct. 2005).

[Nielsen], Jakob Nielsen, "End of Web Design", Alertbox, July 23, 2000.

www.keynote.com

[Bhatti], Nina Bhatti, Anna Bouch, and Allan Kuchinsky, "Integrating User-Perceived Quality into Web Server Design", The 9th International World Wide Web Conference, May 15 – 19, 2000.

www.apdex.org

[Sevcik], Peter Sevcik, "Defining The Application Performance Index", Business Communications Review, (Mar 2005).

[Selvidge], Paula, "How Long Is Too Long To Wait For A Website To Load?", Usability News, (January 1999).

Loosley, Chris, Richard L. Gimarc, Amy C. Spellmann, "E-Commerce Response Time: A Reference Model", Computer Metrics Group 2000.

Norton, Dr. Tim R., "End-To-End Response Time: What to Measure", Computer Metrics Group 1999.