Defining Standards for Web Page Performance in Business Applications

Garret Rempel
MNP Consulting
2500-201 Portage Ave
Winnipeg, MB, Canada R3B 3K6
1-204-924-7503
garret.rempel@mnp.ca

ABSTRACT

Distinctions between standalone computer applications and web-based applications are becoming increasingly blurry, and client-server communication is being used as a part of everyday computing. This is resulting in user expectations of web page performance converging with their expectations of standalone application performance. Existing industry standards for web page performance are widely varied and inconsistent, and standards based on surveying users are especially so. We illustrate this inconsistency with examples from published literature and industry studies.

Considering a specific class of web-based applications (high usage, minimal overhead, business web applications), we attempt to define a set of industry standards by conducting a case study of an implementation of an industry-leading software suite. We measure the application's performance over time and contrast its performance with the frequency of reported end-user performance complaints. Taking these measurements, we define a specific set of measurable performance standards that, when met, would achieve a high level of performance satisfaction among a large percentage of users.

Based on our examination of existing industry standards, we know there are limitations in users' ability to define consistent performance requirements. Here we present a method that proposes to produce a set of performance requirements through a user interview process that closely matches the performance standards defined by the case study. We then examine the results achieved by applying this method to a comparable web application within the same company as the case study to demonstrate that the requirements produced match the performance observations of the case study analysis.

Categories and Subject Descriptors

D.2.1 [**Software Engineering**]: Requirements/Specifications – *Elicitation methods, Methodologies;* D.2.8 [**Software Engineering**]: Metrics – *Performance measures;* H.1.2 [**Models and Principles**]: User/Machine Systems – *Human factors, Software psychology*.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from Permissions@acm.org.
ICPE'15, January 31 - February 4, 2015, Austin, TX, USA.
Copyright is held by the owner/author(s). Publication rights licensed to ACM.
ACM 978-1-4503-3248-4/15/01 …$15.00.
http://dx.doi.org/10.1145/2668930.2688056

General Terms

Measurement, Performance, Human Factors, Standardization

Keywords

Performance; industry standards; requirements; case study; methodology; web applications

1. INTRODUCTION

We are seeing a convergence between standalone applications and the web, which includes traditional and mobile platforms. Cloud computing, network data storage, and sharing information between desktop and mobile devices is becoming ubiquitous and transparent. In many cases it is difficult for a user to know when they are interacting with a piece of technology as to whether it is operating in a standalone manner or if it is communicating with the web [22]. Thus a user's expectations of the performance of their consumer electronics are converging across all platforms and technology. As web browsers and web enabled applications become a dominant platform [25], users expect to be able to do the same things and have the same performance regardless of their technology choices [1]. This convergence in technology means that applying traditional Human-Computer Interface (HCI) standards to web technology is becoming more relevant and important.

Currently there are widely varying opinions on what constitutes an acceptable web page load time. These opinions range from over 30 seconds to less than 1 second, depending on who is espousing the standard and what methodology they have followed to obtain it. There are three main categories of studies in performance literature that establish standards: Physiological Measurements, Empirical Studies, and Surveys.

Physiological Measurements examine physical thresholds or reactions of the human body that are independent of a users' decision-making process. These measurements are effective at providing limits to the level of optimization required to make something 'good enough' and defining thresholds of human perception that, once crossed, further improvement has limited or no additional benefit.

Empirical Studies typically measure the duration of time between the start of a web page request and a point in time when a user chooses to abort the request prematurely (hereafter referred to as "abandonment"). Empirical studies use a variety of tools and testing methodologies to impose known delays and measure when and how many participants abandon their task. They do not directly attempt to measure a participant's emotional state, but rely on measuring concrete actions or decisions that a participant makes as a result of exceeding their wait time tolerance.

Empirical studies are effective at evaluating strategies to overcome or mitigate poor performance by comparing those strategies against a baseline to determine how effective they are at delaying or negating user abandonment.

Surveys typically ask participants to evaluate their emotional reaction, frustration level, or satisfaction level based on page performance as part of an empirical study. They may also ask participants to provide a self-evaluation of how long they will wait before they become frustrated or abandon a web page (their 'wait time tolerance'). The most significant limitation to surveys is that they are subject to a participant's perception and rely on a participant's ability to self-evaluate and quantify an undependable emotional state that can be impacted by many factors [24]. These studies are important however, since "frustration" and "satisfaction" are both emotional reactions that strongly influence user perception of the credibility [6] and quality of a system [3].

1.1 Physiological Measurements

Studies conducted on human interactions with computer systems have given us reasonable upper and lower bounds for several classes of interactions based on their purpose. One of the most widely-cited results is Miller's (1968) powers-of-10 thresholds of 0.1s, 1.0s, and 10s [13] that places limits on a person's perception of instantaneous reaction (0.1s), their continuity of thought (1.0s), and ability to keep their attention focused on the dialogue of interaction (10s) [15].

In terms of conversational interaction with a system, the acknowledgement of "knowledge of results" peaks in effectiveness for simple responses at 0.5s with an upper limit of 2.0s, and delays of more than 4.0s indicate a break in the thread of communication. This is also compatible with the result that distractions can interfere with information being held in short-term memory, and the effects of distractions on short-term memory rapidly increase once a person is aware that they are waiting. This awareness of waiting typically begins to occur in as little as 2 seconds [13].

A further study on gaze fixation demonstrates that a display with multiple independently loading components will result in person fixating on the components that load quickest. A page with a slowly (8s) loading banner advertisement that occupies 23% of the screen space will only receive 1% of a person's attention time, while the same advertisement will receive 20% of a person's attention time when they are only exposed to the screen after rendering has completed [16].

These results indicate the unintended consequences of first-impression response time optimization (as opposed to fully rendered response time optimization). If the most important information is not among the earliest to display onscreen, a person is less likely to give it the attention that it requires. In addition, if unimportant information is rendered earlier than important information, the likelihood increases of becoming distracted and that distraction interfering with their short-term memory and the continuity of their thought process.

1.2 Empirical Studies

The primary focus of this class of study has been to measure the average (mean or median) time that a user will wait while expecting a response before abandoning the process (their 'wait time tolerance'). For systems designed for public consumption, these studies provide an important upper bound on web page response time performance where abandonment is a significant and measurable negative outcome. Several of these studies also evaluate the effectiveness of providing feedback to the user while they are waiting as a tactic for delaying abandonment by communicating with the user that an error has not occurred and something is happening behind the scenes.

These studies have approached the question of abandonment in a variety of methods and have produced a wide range of results and recommendations from less than 2 seconds [14][24] up to 41s [19]. The most constrained result (<2s) is from Shneiderman (1984) and is regarding the pace of human-computer interaction and predates public access to the internet [14]. However, as expectations for offline and online computing converge, this result becomes important to consider as we can observe a steady decrease in the duration of time a person is willing to wait among online-specific studies as shown in Table 1.

Table 1: Wait Time Tolerances for Online-Specific Empirical Studies

Study	Year	Wait Time Tolerance (in seconds)
Bickford [2]	1997	8.5
Ramsay [19]	1998	41
Selvidge* [23]	1999	30
Hoxmeier* [10]	2000	12
Galleta, Henry, McCoy, and Polak* [8]	2002	4-8
Nah [14]	2004	2-4

*These studies included a survey component in an attempt to quantify the impact on user satisfaction levels caused by introducing fixed delays into the rendering process.

The study conducted by Nah (2004) is of particular interest as the conclusions she presents regarding wait time tolerances most closely agrees with Shneiderman's (1984) conclusions regarding human-computer interactions. Specifically, Nah states that "the findings from this study suggest that most users are willing to wait for only about two seconds for simple information retrieval tasks on the Web." [14] She also directly references the results obtained by Shneiderman (1984) and Miller (1968) as comparable findings. However, her conclusions are based upon a method that involves iterative testing and are drawn from her observation that each test subject's wait time tolerance decreases with subsequent failures. Her result is that a person's wait time tolerance converges towards 2s *after experiencing multiple failures* (shown in Table 2).

Table 2: Wait Time Tolerances Measured by Nah (2004)

Without Feedback	Mean	Median	Mode
First Response Failure	13s	9s	5-8s
Second Response Failure	4s	3.6s	2-4s
Third Response Failure	3.3s	2.5s	2-3s
With Feedback	**Mean**	**Median**	**Mode**
First Response Failure	37.6s	22.6s	15-16s, 20-22s, 45-46s
Second Response Failure	17s	8.4s	2-3s
Third Response Failure	6.7s	4.3s	2-3s

Nah's (2004) conclusion that 2s is the 'tolerable' threshold for web page performance is not necessarily supported by the evidence. Her method of providing instantaneous responses to successful page requests interspersed with three pages which fail by never providing a response, she is conditioning her subjects to expect one of two outcomes when clicking a link: instantaneous success, or failure. Therefore she is proving that *when conditioned*, a subject's ability to determine whether or not a page will succeed or fail trends towards 2s by their third failure.

A much more important observation from Nah's (2004) study is when a subject is given a defined task to accomplish (load a web page), 50% of the subjects will abandon that task within 9s (measured by the median wait time tolerance of the First Response Failure, no feedback). It is also telling to observe that poor performance will *decrease* a subject's wait time tolerance. A perceived failure to load a web page due to poor performance will cause a person to more quickly abandon subsequent requests.

For business purposes, the results of these studies are important in establishing a maximum response time limit for web pages to load. They do not however, provide a conclusive target for response times, as typically a business would intend to achieve a much lower abandonment rate than 50%. In order to establish reasonable performance goals, they must be based on people's subjective reactions since a business typically intends to achieve a high rate of satisfaction, and a negligible rate of abandonment due to performance.

1.3 Surveys

Of the three categories of performance studies, surveys provide the most subjective and ambiguous results as they rely on respondents being capable of objectively self-assessing or estimating their own emotional reaction to web page performance. It is also well-established that a person's threshold for frustration and wait time tolerance are impacted by many variables [24]. Methodology also plays an important part of assessing survey results, as some surveys that use a multiple-choice structure may be subject to Central Tendency Bias or Position Bias, wherein respondents will prefer options presented in the middle or at the beginning of a list unless they hold very strong opinions on the question being presented [9].

Notable results from surveys on emotional reactions that are conducted within an empirical study framework result in recommendations that range from 30s [23], to 12s [10], to 4-8s [8]. Results from non-empirical standalone surveys include 8s [27] (which provides no justification for using the threshold of 8s), 4s [11], and 2s [7]. The conclusions drawn from these surveys, much like the empirical studies, are based upon average (mean or median) answers provided by the respondents [20]. Comparing conclusions, we can also see an unusually rapid drop in performance expectations by respondents, a drop that does not appear to correspond with an equivalent improvement in consumer connectivity. In regards to the JupiterResearch (2006) and Forrester (2009) surveys, the results as given show a drop of 50% in performance expectations (from 4s to 2s). However, broadband penetration in the United States only increased from 20.3% to 25.5% between Q4 2006 and Q4 2009 [17] and 16.9% to 23.1% for the entire OECD survey region. Since expectations of web page performance are primarily influenced by experience, it is unlikely that the expectation of web page response times would drop so significantly on average over the same timeframe. This is also supported by research that shows that the average web page size has continuously increased since 2003, and from 312KB to 507KB (62.5%) between 2008 and 2009 alone [26]. This increase continues to appear until we see an average size of 1510KB in 2014 [5]. This increase in web page size is only partially offset by improving connectivity and has resulted in a net *increase* in typical web page loading times over that time period.

Re-examining the results published by JupiterResearch (2006) and Forrester (2009) by examining their high percentile responses and combining those results along with an additional survey conducted by Rempel (2014), we can see a significantly more consistent and useful result. Combining the three surveys produces a conclusion that at least 90% of respondents would be satisfied with response times of 1s or less, and at least 80% of respondents would be satisfied with response times of 2s or less. Additionally, on average we would expect a satisfaction level of 99% with response times of 1s or less, and 90% with response times of 2s or less [20].

1.4 Findings

Studying the subject of industry standards as they apply to web page performance, it is apparent that there is a lack of consistency in the way the topic is studied, measured, and the conclusions that are drawn. There are however, a few observations that can be made that are useful from a business perspective.

As traditional applications, the web, and mobile devices converge, the expectations that people have in terms of web performance will converge with their expectations of traditional human-computer interfaces. Therefore it is reasonable to consider that original HCI research by Miller (1968) can serve as a gold standard for performance targets, even if they are not widely achievable.

- The threshold for instantaneous feedback is <0.1s.
- The threshold for train of thought (typical web request-response interaction) is <1.0s.
- The threshold for complex operation is <10.0s.

In addition, wait time tolerance improves when dynamic feedback is presented. There is a case to be made for using 2s as a threshold as referenced by Miller (1968), Shneiderman (1984), Nah (2004), and Forrester (2009) with varying levels of support. It is reasonable to suggest that 2s be used as an upper bound for basic, non-instantaneous transactions, and that any transactions that consistently take longer than 2s should take advantage of feedback techniques to avoid user abandonment. Operations that are known to exceed 10s should include additional feedback or pre-operation estimates of expected wait time to improve a person's engagement with the system and ease potential frustration.

It is also important to note that *consistency* in performance is important, and that page interactions should not deviate outside of 25% to 200% of the mean [24] to avoid anxiety or frustration caused by unexpected or unusual variances in response time.

2. CASE STUDY

This section describes a study that was performed on an internal business web application in production in order to examine user feedback (complaints) about system performance as compared to actual performance metrics gathered on the system. This study differs from the previously described studies in that it is a passive, unsolicited study of actual user behavior and satisfaction levels in

a real life situation, instead of a study comprised of volunteers or survey respondents.

2.1 System Under Scrutiny

The system being studied is the primary client information tracking and incident reporting system of an international company and is recognized as an industry-leading platform, supported by a reputable international vendor. The system is used by 1,200 users in offices spanning 5 time zones, from the east coast to the west coast of Canada and the United States. Many of the system's users are required to use this application as part of their primary duties, often while actively communicating with their customers, while using the application to perform data entry that their customers are providing over the phone or in person. During the course of a business day, peak usage has been measured at 800 simultaneous logins and 50,000 page requests per hour over a four-hour window. On an average weekday, the system receives 440,000 page requests and peaks at 510,000 page requests on the busiest day of the week. The system also receives approximately 10,000,000 page requests per month.

This web application has several advantages as a subject for study:

- The period of time under scrutiny includes initial go-live, as well as several months of stabilization in which systemic performance problems were observed.
- The system's users are motivated and encouraged to report impressions of poor performance via a centralized service desk.
- Users do not have any recourse to abandon the application as no alternatives exist to provide workarounds to problems they encounter.

As a business application, the system has been designed as a high-efficiency system. All secondary requests to acquire resources from a web page are made to the same server that the web application is hosted on, eliminating the need for secondary DNS lookups or additional connections to be established. Resources in use are also very minimal and are cached by the browser after the initial request. Page structures are also very simple and there are no javascript commands issued with the onload directive, resulting in a very fast render time.

2.2 Methodology

The time period of this study is January 2012 through May 2012, as well as October 2013, all of which encompasses:

- Initial go-live of the system.
- A period of systemic performance problems.
- Incremental improvements in system performance as patches are released and performance defects are resolved.
- A comparative period (October 2013) in which the system is stable, performance is optimized, and no further patches are being developed for the system.

In order to establish an upper bound for server communication time, a load generation tool was used to generate a known sequence of requests to an identical production support server using a fixed interval between the end of one request and the beginning of the next over a period of four hours. This process was conducted from an agent local to the system's host server on the same local subnet, and repeated from an agent deployed in an office in the furthest geographic location from the host server (Southern California). The number of completed requests was compared between the two runs to determine an upper bound for communication latency between the two locations. This test sequence was repeated at different times of the day and days of the week and the results averaged to produce an expected communication latency margin for each request.

The developer tools in Google's Chrome browser were used to obtain a random sampling of requests and analyze the duration of time required to parse and render the final result of the page. These timings were weighted and averaged to determine the expected render time for each request.

The production server was instrumented to record and log the interval between receipt of each request and transmission of the final response. These logs were aggregated and analyzed across every request made to the server to generate a count of requests by hour and by day into 0.5s response time buckets.

Service desk logs were examined to obtain the number of issue tickets opened (complaints) related to perceived performance problems with the web application and correlated with the measured performance of the web application.

2.3 Findings

2.3.1 Communication Latency

To calculate the margin of communication latency that exists for remote locations, we executed the same test plan over a fixed duration from an agent local to the host server and from one that is remote. The request frequency was then analyzed to calculate the margin that existed between the two tests. This process was performed three times at different times of the day and different days of the week to reduce the impact of network traffic on the results. The following calculation was used to calculate the communication latency.

Let r be the number of completed requests, t be the total time of the test in seconds, and a be the average time between requests (which is the sum of the fixed think time plus the measured average response time of all requests). Let R represent the test execution from the remote agent, and L represent the test execution from the local agent. The communication latency at the remote agent is defined by:

$$M = a_L - [(r_R \cdot t_L \cdot a_L) / (r_L \cdot t_R)]$$

This value is averaged over multiple iterations, producing a final calculated average communication latency of 0.481 seconds.

2.3.2 Render Time

To calculate an estimated render time margin, Google's Chrome browser developer tools were used to analyze the network timeline of a series of page request / responses. Prior to the test, the browser's cache was cleared in order to evaluate the difference in time between an initial request to the server, and a request where resources had already been cached.

Prior to resource caching being completed, secondary resource requests and rendering time required up to 0.190s to process between receipt of the initial server response and the final load event being fired. All subsequent requests (once the resource cache was established) required between 0.004s to 0.047s to complete rendering. This resulted in an effective average of 0.026s.

2.3.3 *Aggregated Response Percentiles*

The aggregated server processing times provided by server instrumentation have been adjusted by adding a 0.5s margin to account for the measured average communication latency and rendering and the result is shown in Table 3. The values presented in the table are the percent of requests during each month that completed within the specified time constraint. Major percentile thresholds are color-coded so that measurements that fall below the 85th percentile are red, below the 90th percentile are orange, below the 95th percentile are yellow, and 95th percentile and above are green.

Table 3: Server Processing Percentiles with Rendering and Latency Margins

	Percent of Requests Completed within Range					
	Jan 2012	**Feb 2012**	**Mar 2012**	**Apr 2012**	**May 2012**	**Oct 2013**
< 1.0 s	64.43	63.82	59.98	67.87	69.06	70.22
< 1.5 s	80.65	79.62	76.18	81.57	83.76	87.27
< 2.0 s	87.47	86.18	82.83	87.17	89.62	92.17
< 2.5 s	91.40	90.40	87.32	91.01	93.37	97.81
< 3.0 s	93.64	92.72	89.94	92.76	95.27	
< 3.5 s	94.94	94.12	91.65	93.82	96.29	
< 4.0 s	95.82	95.09	92.94	94.82	97.00	
< 4.5 s	96.51	95.88	94.05	96.01	97.66	
< 5.0 s	97.11	96.57	95.02	96.83	98.23	

From this information we can observe four distinct phases of application performance: Go-live (Jan / Feb 2012), Decay (Mar 2012), Optimization (Apr / May 2012), and Stable (Oct 2013). These four periods are mirrored by the level of complaints issued to the service desk over perceived performance problems during the same timeframes as shown in Table 4.

Table 4: Performance Complaints to Service Desk

	Jan 2012	**Feb 2012**	**Mar 2012**	**Apr 2012**	**May 2012**	**Oct 2013**
Complaints*	17	20	22	21	13	0

*Jan-May monthly values are estimated based on reported number of complaints received/week.

There is a clear connection between the number of service desk complaints issued about performance and the aggregated web page performance percentiles of the system. As performance decays (illustrated by major percentile thresholds dropping into lower response time buckets), the number of complaints increases. Conversely as the performance level of the system increases, the number of complaints drops until a certain level is reached at which point no new performance complaints are forthcoming.

Reorienting the results from Table 4 in order to show response time levels for each major percentile, Table 5 and Figure 1 compare the four phases of application performance and how they changed over time.

Although the number of actual performance complaints made to the service desk appears small, the number of unhappy users is typically much larger than the number that will issue a formal complaint. In a public consumer system, only 1 out of 26 unhappy customers will lodge a formal complaint, while many of the remaining customers will simply never return [4][18]. Because of the nature of this system being an internal business system, it is possible that the number of unhappy users who do not complain is smaller. However, regardless of the actual factor involved, the number of complaints is a good indicator of user satisfaction.

Table 5: Response Time Levels for Major Percentiles by Phase

Percentile	**Go-live (Jan/Feb 2012)**	**Decay (Mar 2012)**	**Optimization (Apr/May 2012)**	**Stable (Oct 2013)**
80th	<1.5s	<2.0s	<1.5s	<1.5s
85th	<2.0s	<2.5s	<2.0s	<1.5s
90th	<2.5s	<3.5s	<2.5s	<2.0s
95th	<4.0s	<5.0s	<3.0s/<4.5s	<2.5s
Complaints per Month	*18.5*	*22*	*17*	*0*

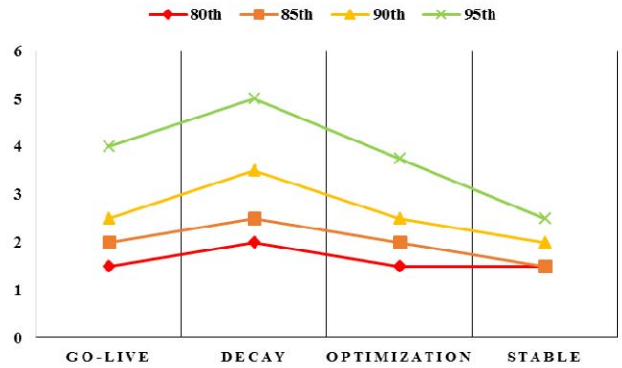

Figure 1. Response Time Levels for Major Percentiles by Phase

Based on this study we have sufficient information to set performance goals for future systems that may be implemented at this corporation as follows:

- 95% of all page requests must be completed within 2.5s.
- 90% of all page requests must be completed within 2.0s.
- 85% of all page requests must be completed within 1.5s.

3. GATHERING PERFORMANCE REQUIREMENTS PROCESS

Given the wide discrepancy in web page performance standards that use a survey as part of its data gathering process, and the wide spread in responses (0.5s to 60s) given when respondents were asked to provide their own estimate for their wait time tolerance [20], it is uncertain whether performance requirements can be reasonably obtained through an interview process with business users for a new web application. If two business users were interviewed independently for performance requirements without guidance, it is entirely possible that their response time targets could range from <0.5s to <30s. This section describes an effort to refine the process of establishing performance requirements by presenting guidelines to business users prior to

soliciting their input, in an effort to create a set of performance requirements that closely match the results generated by the case study. To match the case study results, on aggregate we would need to see 95% of all web page requests achieve an end-to-end response time of 2.5s or less. To meet this target and to stay in line with prior studies of a <2s standard, we would require a majority of page performance targets be set at <2s. A limited number of pages may have larger performance targets depending on function and still meet these guidelines. Pages with larger performance targets can also be candidates for alternate feedback strategies to improve end user perception [2][14].

3.1 Response Time Performance Categories

In this attempt to establish performance requirements for a new application, a set of performance requirement categories were established with assistance from system developers to define examples. Each performance requirement category (shown in Table 6) had four components, a definition (with examples), a target response time level, a maximum response time level, and a stability measurement (a percentile value to be used to measure against the target). Percentiles are used to measure web page performance instead of an average response time in order to provide a better bound on the page performance and smooth out any spikes in measured response times [12][21]. It is also used to provide a comparable metric to our server processing time measurements as defined in the prior case study.

Table 6: Performance Requirement Categories

Category Name	Target Response Time	Maximum Response Time	Stability (Percentile)
Basic Operations Ex. Most Standard Pages or Simple Operations	<2 s	<2 s	95th
Complex or Ambiguous Search or Save Operations Ex. Major Save Operations, Large Result Set Searches	<5 s	<5 s	90th
Integration or Major Calculation Operations Ex. Upload Documents, Synchronous Interfaces, Complex Calculations	<5 s	<15 s	85th
Heavyweight Operations Extremely Complex Calculation and Data Processing Operations, Resource Intensive Interfaces	<10 s	<30 s	85th

To evaluate the performance of a specific web page or operation after a test cycle, its measured response time performance is compared to the target and maximum response times defined in its category. A page is considered to have passed under a typical load when:

- Its percentile response time measurement meets or betters the target response time.
- Its overall maximum response time measurement meets or betters the maximum response time.

A page is considered to have passed under maximum peak load when:

- Its percentile response time measurement meets or betters the maximum response time.

3.2 Requirement Definition Process

In the example presented in the previous case study, the functional requirements of the system defined 259 distinct web pages or operations in the system that could be evaluated for performance. The performance requirement categories were presented to the business users and they were then given instructions to categorize each page into one of the pre-defined performance categories whenever possible. They were also instructed that if there was sufficient reason that a page could not be placed into one of the categories, the business could create new performance requirements for that page. At the end of the process, the pages were categorized by the business as shown in Table 7.

Table 7: Business User Allocation of Web Pages into Response Time Categories

Category Name	# of Pages	% of Total Pages
Basic Operations	222	85.71
Complex or Ambiguous Search or Save Operations	29	11.20
Integration or Major Calculation Operations	1	0.39
Heavyweight Operations	7	2.70

Calculating the weighted average performance target response time value for all of the pages in the system produced a value of 2.56s, and the weighted average performance maximum response time value was 3.14s.

During the next stage of the requirements process, the business users were asked to weight the web pages and operations based on number of requests per day. These weightings were used during a performance test cycle to simulate production workload. The test results were analyzed based on the number of page requests made to each of the pages and the results are shown in Table 8.

Table 8: Performance Testing Page Request Frequency by Response Time Category

Category Name	# of Page Requests During Test Cycle	% of Total Page Requests
Basic Operations	353,737	89.54
Complex or Ambiguous Search or Save Operations	33,550	8.49
Integration or Major	2,942	0.74

Calculation Operations		
Heavyweight Operations	4,819	1.21

Performing a similar calculation as was done for the number of web pages, based on the number of actual page requests, the weighted average performance target response time value for all of the pages in the system was 2.37s, and the weighted average performance maximum load time value was 2.69s.

If achieved in a production environment, the requirements presented here closely correspond to the desired performance target of achieving 95% of all web page requests in under 2.5s, and matches the calculated wait time tolerance level of the users of this type of system, in this particular company, as illustrated in the above case study.

4. CONCLUSIONS

Industry performance standards are widely variable and inconsistently structured and researched. However, a careful study of a web application that exists in a controlled environment shows that the actual wait time tolerance of the users in the study closely aligns with the most popular performance recommendations of <2s.

By using this case study to pre-define performance target categories with assistance from business analysts and system developers, business users with no particular training or experience with performance requirements were able to independently define performance requirements that closely aligned with the observed optimal performance state of an existing production application.

5. ACKNOLWEDGEMENTS

The author has benefited from discussions with Glenn Hemming.

6. REFERENCES

[1] Anderson, J. 2012. The Web Is Dead? No. Experts expect apps and the Web to converge in the cloud; but many worry that simplicity for users will come at a price. *Pew Research Center's Internet & American Life Project.* Retrieved August 8, 2014, from Pew Research Center: http://www.pewinternet.org/files/old-media/Files/Reports/2012/PIP_Future_of_Apps_and_Web.pdf

[2] Bickford, P. 1997. Worth the Wait. *Netscape's Developer Edge, Netscape Communications (online)*, Mountain View, CA, USA (1997). Retrieved August 8, 2014, from Archive.org: http://web.archive.org/web/20040913083444/http://developer.netscape.com/viewsource/bickford_wait.htm

[3] Bouch, A., Kuchinsky, A., and Bhatti, N. 2000. Quality is in the eye of the beholder: meeting users' requirements for Internet quality of service. In *Proceedings of the SIGCHI conference on Human Factors in Computing Systems* (CHI '00). ACM, New York, NY, USA, 297-304. DOI=10.1145/332040.332447 http://doi.acm.org/10.1145/332040.332447c

[4] Digby, J. 2010. 50 Facts about Customer Experience. *Return on Behavior Magazine (online).* Retrieved August 8, 2014, from Archive.org: https://web.archive.org/web/20140210221957/http://returnonbehavior.com/2010/10/50-facts-about-customer-experience-for-2011/

[5] Everts, T. 2014. State of the Union, Ecommerce Page Speed & Web Performance. Retrieved August 8, 2014, from Radware: http://www.webperformancetoday.com/2014/04/29/spring-2014-state-union-ecommerce-page-speed-web-performance-infographic/

[6] Fogg, B.J., Marshall, J., Laraki, O., Osipovich, A., Varma, C., Fang, N., Paul, J., Rangnekar, A., Shon, J., Swani, P., and Treinen, M. 2001. What makes Web sites credible?: a report on a large quantitative study. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (CHI '01). ACM, New York, NY, USA, 61-68. DOI=10.1145/365024.365037 http://doi.acm.org/10.1145/365024.365037

[7] Forrester Consulting. 2009. eCommerce Web Site Performance Today; An Updated Look At Consumer Reaction To A Poor Online Shopping Experience. Retrieved August 8, 2014, from Damco Group: http://www.damcogroup.com/white-papers/ecommerce_website_perf_wp.pdf

[8] Galletta, D.F., Henry, R., McCoy, S. and Polak, P. 2004. Web Site Delays: How Tolerant are Users?, *Journal of the Association for Information Systems*: Vol. 5: Iss. 1, Article 1. http://aisel.aisnet.org/jais/vol5/iss1/1

[9] Gingery, T. 2009. Survey Research Definitions: Central Tendency Bias. Retrieved August 8, 2014, from Cvent: http://survey.cvent.com/blog/market-research-design-tips-2/survey-research-definitions-central-tendency-bias

[10] Hoxmeier, J.A. and DiCesare, C. 2000. System response time and user satisfaction: an experimental study of browser-based applications. *Proceedings of the Americas Conference on Information Systems*, (Long Beach, CA, USA, 2000), Association for Information Systems, 140-145. http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.99.2770

[11] JupiterResearch 2006. Retail Web Site Performance; Consumer Reaction to a Poor Online Shopping Experience. Retrieved August 8, 2014, from Akamai: http://www.akamai.com/dl/reports/Site_Abandonment_Final_Report.pdf

[12] Meier, J.D., Farre, C., Bansode, P., Barber, S. and Rea, D. 2007. Performance Testing Guidance for Web Applications, *Microsoft Patterns & Practices (Chapter 15 – Key Mathematic Principles for Performance Testers).* Retrieved August 8, 2014, from MSDN: http://msdn.microsoft.com/en-us/library/bb924370.aspx

[13] Miller, R.B. 1968. Response time in man-computer conversational transactions. In *Proceedings of the December 9-11, 1968, fall joint computer conference, part I* (AFIPS '68 (Fall, part I)). ACM, New York, NY, USA, 267-277. DOI=10.1145/1476589.1476628 http://doi.acm.org/10.1145/1476589.1476628

[14] Nah, F. 2004. A Study on Tolerable Waiting Time: How Long Are Web Users Willing to Wait?, *Behaviour and Information Technology*, (Lincoln, NE, USA, 2004), Vol. 23, No. 3 (May 2004). Retrieved August 8, 2014 from

University of Nebraska-Lincoln: http://cba.unl.edu/research/articles/548/download.pdf

[15] Nielsen, J. *Usability Engineering*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1993.

[16] Nielsen, J. 2010. Website Response Times. Retrieved August 8, 2014, from Nielsen Norman Group: http://www.nngroup.com/articles/website-response-times/

[17] OECD 2013. Historical penetration rates, Fixed and Wireless broadband, G7 (June 2013). Retrieved August 8, 2014, from OECD Broadband Portal: http://www.oecd.org/sti/broadband/1i-BBPenetrationHistorical-G7-2013-06.xls

[18] Quinn, P. 2014. The Customer Complaint Iceberg. Retrieved August 8, 2014 from PeoplePulse: http://www.peoplepulse.com.au/Customer-Experience.htm

[19] Ramsay, J., Barbesi, A. and Preece, J. 1998. A psychological investigation of long retrieval times on the world wide web, *Interacting with Computers* (1998), 10(1), 77-86. DOI=10.1016/S0953-5438(97)00019-2 http://iwc.oxfordjournals.org/content/10/1/77.abstract

[20] Rempel, G. 2014. Web Performance Standards: Finding Value in User Surveys. Retrieved August 8, 2014, from Mincing Thoughts: http://mincingthoughts.blogspot.com/2014/07/web-performance-standards-finding-value.html

[21] Rhea, R. 2012. Performance Test Best Practices With Rational Performance Tester. Retrieved August 8, 2014, from IBM DeveloperWorks: https://www.ibm.com/developerworks/community/groups/service/html/communityview?communityUuid=a9ba1efe-b731-4317-9724-a181d6155e3a#fullpageWidgetId=W5f281fe58c09_49c7_9fa4_e094f86b7e98&file=b3e1526b-8981-4e42-826d-d8eadc569a13

[22] Schindler, E. 2007. The Convergence of Desktop, Web and Mobile Clients. Retrieved August 8, 2014, from CIO: http://www.cio.com/article/2437560/developer/the-convergence-of-desktop--web-and-mobile-clients.html

[23] Selvidge, P. 1999. How long is too long for a website to load?. *Usability News*, 1(2). Retrieved August 8, 2014, from Archive.org: http://web.archive.org/web/20020404143111/http://psychology.wichita.edu/surl/usabilitynews/1s/time_delay.htm

[24] Shneiderman, B. 1984. Response time and display rate in human performance with computers. *ACM Comput. Surv.* 16, 3 (September 1984), 265-285. DOI=10.1145/2514.2517 http://doi.acm.org/10.1145/2514.2517

[25] Wang, H., Moshchuk, A. and Bush, A. 2009. Convergence of Desktop and Web Applications on a Multi-Service OS. In *HotSec '09*, (Montreal, CA, 2009). https://www.usenix.org/legacy/event/hotsec09/tech/full_papers/wang.pdf

[26] Web Site Optimization 2014. Average Web Breaks 1600K. Retrieved August 8, 2014, from WebSiteOptimization: http://www.websiteoptimization.com/speed/tweak/average-web-page/

[27] Zona Research Inc. 1999. The Economic Impacts of Unacceptable Web-Site Download Speeds. Retrieved August 8, 2014, from WebPerf: http://www.webperf.net/info/wp_downloadspeed.pdf