

Bachelor-Thesis

Web Performance für den mobilen Endanwender

Zusammenfassung

2. März 2015

Columbus Interactive
Eywiesenstraße 6
88212 Ravensburg

Fakultät Elektrotechnik und Informatik
Studiengang Angewandte Informatik
Hochschule Ravensburg-Weingarten
Doggenriedstraße, 88250 Weingarten

Autor:
Andreas Lorer
`andreas.lorer@hs-weingarten.de`
88250 Weingarten
Wilhelmstraße 4



Bachelor-Thesis

Web Performance für den mobilen
Endanwender

Inhaltsverzeichnis

1	Einleitung	2
2	Motivation	2
2.1	Zielsetzung	2
3	Eigene Leistung	3
4	Ist-Zustand	4
5	Grundlagen	4
5.1	Begriffe	4
5.1.1	Pattern und Antipattern	4
5.1.2	HTTP-Request Komponenten	4
5.1.3	Round Trip Time (RTT)	5
5.1.4	TCP Slow Start	5
5.1.5	Above The Fold	6
5.1.6	Kritischer Rendering-Pfad	7
5.2	Perceived Performance	9
5.3	Latenz	10
6	Entwicklung	10
6.1	Tools	10
6.1.1	Google Chrome Developer Tool	10
6.1.2	Webpagetest	10
6.1.3	Google Spreadsheet	10
6.2	Ausgangspunkt	10
6.3	Prozess der Suche	10
6.4	Prozess der Validierung	10
7	Best-Practices	10
7.1	Serverseitig	10
7.1.1	Verringern von DNS Lookups	10
7.1.2	HTTP Requests	10
7.1.3	Caching	10
7.1.4	Pagespeed Mod	10
7.1.5	Images	10
7.2	Clientseitig	11
8	Workflow	11
8.1	Performanten Code schreiben	11
8.2	Minify und Uglify	11
8.3	Concatenating	11
8.4	Task Manager	11
8.5	Dependency Manager	11
8.6	Generators	11

1 Einleitung

2 Motivation

Larry Page, CEO und Mitgründer von Google, sagt:

„As a product manager you should know that speed is the number one feature.“ (Holzle 2010)

Niemand mag es zu warten, auch nicht auf eine Website. Die Studie „The Psychology of Web Performance“ kam bereits im Jahr 2008 schon auf folgende Ergebnisse:

„Slow web pages lower perceived credibility and quality. Keep your page load times below tolerable attention thresholds, and users will experience less frustration, lower blood pressure, deeper flow states, higher conversion rates, and lower bailout rates. Faster websites are actually perceived to be more interesting and attractive.“ (WebSiteOptimization.com 2008)

Das haupt Vermarktungsargument für den Chrome Browser war damals: er sei schneller als die Konkurrenz. Tatsächlich ist für Google Geschwindigkeit alles. Deshalb hat Google im Jahr 2010 angekündigt, dass Geschwindigkeit in die Berechnung des **Page Rankings** mit einfließt.

„Faster sites create happy users and we’ve seen in our internal studies that when a site responds slowly, visitors spend less time there. [...] Recent data shows that improving site speed also reduces operating costs. Like us, our users place a lot of value in speed — that’s why we’ve decided to take site speed into account in our search rankings“ (Google 2010)

Aktuell (2015) geht Google sogar noch einen Schritt weiter und informiert tausende Webmaster per E-Mail über die schlechte Usability ihrer Websites für mobile Besucher und warnt ausdrücklich vor dementsprechend „angepassten Rankings“. (t3n 2015) Im Hinblick auf die Zukunft wird der Marktanteil an mobilen Internetnutzern noch weiter wachsen und die Optimierung der Ladezeiten gewinnt dadurch noch mehr an Bedeutung. Zwischen 2011 und 2014 stieg die Anzahl der Smartphone Nutzer von 18% auf 50% an. Dies ist ein Wachstum von 32% innerhalb von nur 3 Jahren. (TNS Infratest 2014)

Die Antwort auf diesen Trend läutete eine Ära ein, die wir heute unter dem Namen **Responsive Webdesign** kennen. „Responsive“ muss aber sehr viel mehr bedeuten, als nur eine angepasste Darstellung für eine bestimmte Art von Gerät. „Two out of three mobile shoppers expect pages to load in 4 seconds or less.“ (Radware 2013). Der Anwender erwartet also auf dem Smartphone ähnliche oder gleiche Ladezeiten wie er auch von der Nutzung eines Desktop-Pc’s gewohnt ist. Diese Erwartungen werden von dem Großteil der im Internet besuchbaren Seiten nicht erfüllt. Der Inhalt einer Seite muss darum so aufbereitet werden, dass dieser auch auf Geräten mit langsamer Internetverbindung, hoher Latenz und einem begrenzten Datentarif, in einer für den Anwender annehmbaren Geschwindigkeit, angezeigt werden kann.

2.1 Zielsetzung

Um gängige Methoden und Techniken der Ladezeit Optimierung anzuwenden wird das Projekt anhand der Website <http://andreaslorer.de> durchgeführt. Das Ziel ist es, die Ladezeit der

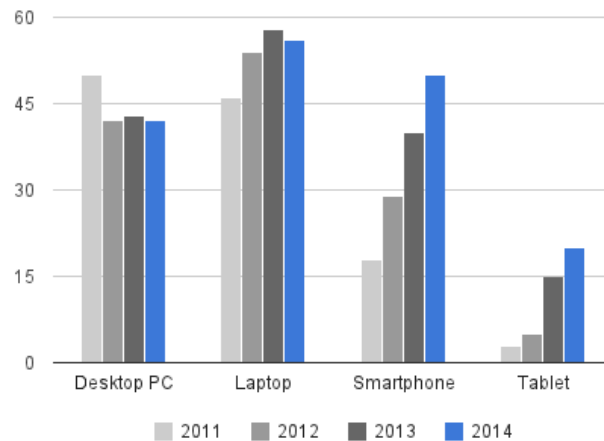


Abbildung 1: Gerätenutzung in der Gesamtbevölkerung (2011 – 2014)(TNS Infratest 2014)

Website auf dem Smartphone, als auch auf dem Desktop von 10 Sekunden auf unter 1 Sekunden zu verringern. Mit Ladezeit ist dabei nicht die Zeit gemeint, die benötigt wird um die Website komplett zu laden, sondern die Zeit bis ein erste visuelle Rückmeldung für den Anwender zu sehen ist. Diese vom Anwender wahrgenommene Rückmeldung nennt man auch „Perceived Performance“ und bedeutet, dass die Ladezeit als schneller empfunden wird, als es eigentlich laut Messwerten der Fall ist. Näheres dazu wird in Punkt 5.2 beschrieben.

3 Eigene Leistung

Meine Leistung besteht darin, einen Leitfaden zu erstellen, der einen Gesamtüberblick ermöglicht. Die Arbeit soll es dem Leser ermöglichen Fehler in der Struktur von Webanwendungen zu finden, die für die Geschwindigkeit hinderlich sind. Es soll herausgefunden werden, was die „Best Practices“ sind um die Ladezeit zu minimieren, wie ein moderner „workflow“ aussehen kann, damit eine Webanwendung schon bei seiner Entstehung schnell ladet und im Projektverlauf schnell bleibt. Des weiteren soll erklärt werden, was für Herausforderungen es zu meistern gilt um eine schnelle Webanwendung zu erreichen, welche Tools es gibt und welche Vor- oder Nachteile diese mit sich bringen.

Diese Arbeit befasst sich nicht, mit der Geschwindigkeit von Datenbanken, SQL-Abfragen oder sonstigen Problemen die durch einen Engpass ein schnelles Laden der Seite verhindern könnten.

4 Ist-Zustand

Die Webseite ist auf einem **shared Hosting**¹ aufgesetzt und antwortet auf ein Ping Kommando in rund 13ms. Dadurch, dass es keine Möglichkeit gibt **root Rechte**² auf einem shared hosting zu bekommen können so manche serverseitige Einstellungen nicht durchgeführt werden. Diese werden dann zwar Aufgezeigt, aber kommen für dieses Projekt nicht zum Einsatz.

Die Website hat als Ausgangsbasis einen gängigen Aufbau. Sie besteht aus einer Bilder Galerie basierend auf PHP und dem Bootstrap Framework.

5 Grundlagen

5.1 Begriffe

5.1.1 Pattern und Antipattern

„Entwurfsmuster (englisch design patterns) sind bewährte Lösungsschablonen für wiederkehrende Entwurfsprobleme sowohl in der Architektur als auch in der Softwarearchitektur und -entwicklung. Sie stellen damit eine wiederverwendbare Vorlage zur Problemlösung dar, die in einem bestimmten Zusammenhang einsetzbar ist.“ (Wikipedia 2015)

„Während „Design Patterns“ in der Software-Entwicklung allgemein übliche und bekannte Ansätze sind, um Probleme zu lösen, sind Anti-Patterns Negativ-Beispiele – die zeigen, wie man es nicht macht - von bereits durchgeführten, gescheiterten Projekten, die dem erkennenden Mitarbeiter zielgerichtete Hinweise darauf geben, wie die Aufgabenstellung besser gelöst werden könnte. Als Synonym ist auch der Begriff Negativmuster im Gebrauch. Es ist tatsächlich möglich, daß das, was gestern noch als allgemein gangbarer Lösungsweg bezeichnet wurde, heute schon ein „Antipattern“ ist [...]“ (Stepken 2006)

5.1.2 HTTP-Request Komponenten

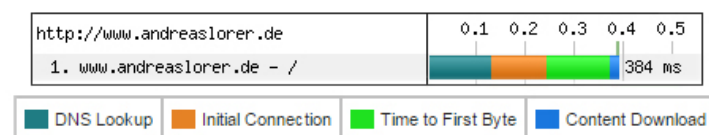


Abbildung 2: Anfrage der HTML-Datei vom server <http://andreaslorer.de> (Abbildung nach <http://webpagetest.org>)

- **DNS Lookup:** Um eine Verbindung mit dem Server herzustellen benötigt das HTTP Protokoll die IP Adresse des Ziels. Das heißt beim DNS Server wird für den Namen „<http://andreaslorer.de>“ die zu diesem Namen zugehörige IP Adresse zurückgegeben.

¹Bei shared Hosting werden mehrere Websites von verschiedenen Website-Betreibern von dem gleichen Webserver gehostet. Bei Shared Hosting teilen sich in der Regel Hunderte andere Websites einen Server (ItWissen.info 2015)

²Standardmäßig existiert unter Linux immer ein Konto für den Benutzer „root“ mit der User-ID 0. Dies ist ein Systemaccount mit vollem Zugriff auf das gesamte System, und damit auch auf alle Dateien und Einstellungen aller Benutzer (wiki.ubuntuusers 2014)

- Initial Connection bezeichnet die Zeit die vergeht, bis eine Verbindung zum Server hergestellt wurde und eine Kommunikation zwischen Browser und Server stattfinden kann. Hierbei findet der sogenannte TCP „Three-Way-Handshake“ statt.³

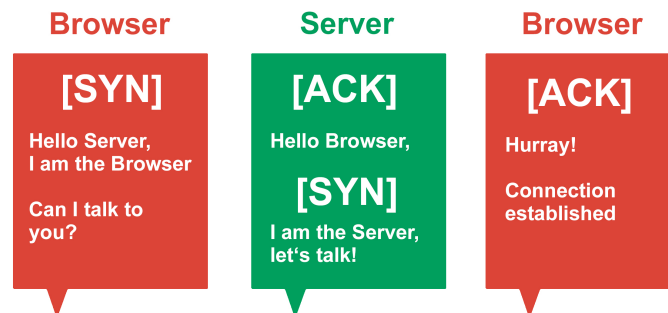


Abbildung 3: Three-Way-Handshake zum Aufbau einer TCP Verbindung zwischen Browser und Server (Eigene Abbildung nach: (Stefanov 2011))

- TTFB: Ist die Abkürzung für „Time to first byte“. Dieser Begriff beschreibt die Zeit die vergeht, bis das erste Byte vom Server beim Browser ankommt. Die Geschwindigkeit und Güte des Servers, Zugriffszeiten auf eine Datenbank oder die Entfernung des Servers können diesen Wert hauptsächlich beeinflussen.
- Content Download: Die Zeit die benötigt wird bis die Datei vom Server Heruntergeladen wurde.

Wie in Abbildung 2 zu sehen ist, ist die Zeit für das Herunterladen des HTML Dokuments verschwindend gering im Vergleich zu der Zeit die damit verbracht wurde, um das Herunterladen zu ermöglichen. Dieses Problem ist vor allem auf Smartphones ausschlaggebend dafür, dass Webanwendungen sehr viel langsamer laden als mit einer Kabelverbindung. Das Thema wird in Punkt 5.3 noch ausführlicher behandelt.

5.1.3 Round Trip Time (RTT)

„Round Trip Time“ wird im Deutschen Paketumlaufzeit genannt. Es bezeichnet die Zeit die ein Datenpaket braucht um in einem Netzwerk von Sender A zu Empfänger B und wieder zurück zu gelangen.

5.1.4 TCP Slow Start

Ein Round Trip kann nicht beliebig viele Bytes transportieren sondern ist durch die sogenannte „Congestion Window Size“⁴ limitiert. Der Überbegriff für dieses Verhalten nennt sich „Slow Start“

- Congestion Control: Nach dem eine neue Verbindung per TCP aufgebaut wurde, können weder Server noch Client wissen, wie schnell die Verfügbare Bandbreite ist, mit der Daten ausgetauscht werden können. Um das Netzwerk, vor einem Datenstau zu schützen, wird mit einem sehr niedrigen Wert begonnen, der dann ansteigt bis das Limit erreicht ist. Dieses Verhalten nennt sich auch „Congestion Control“ und verhindert das Aufstauen von Daten.

³TCP ist das meistgenutzte Verbindungsprotokoll im Internet. Für ein tieferes Verständnis empfiehlt sich dieser Artikel: [High Performance Browser Networking - Chapter 2: Building Blocks of TCP](#)

⁴engl. congestion: Stauung, Überlastung, Anhäufung

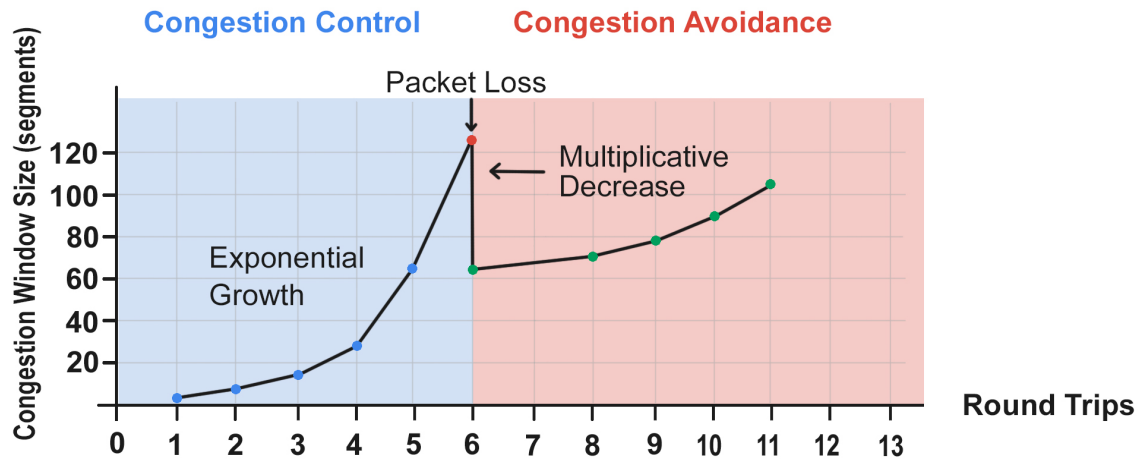


Abbildung 4: Congestion Control und Congestion Avoidance (Eigene Abbildung nach (Grigorik 2013))

- Congestion Window Size: Diese Größe bestimmt, wieviel Bytes der Server pro Segmente schicken darf, bis diese vom Client per ACK (acknowledgement) bestätigt werden müssen. Die Größe der Segmente ist Standardmäßig 1460 bytes und die Rate bis zum ACK ist im April 2013 von 4 auf 10 Segmente erhöht worden. (Grigorik 2013). In der Grafik wird davon ausgegangen, dass der erste Round Trip 4 Segmente senden darf. Die Datenrate wächst exponentiell an, damit möglichst schnell die volle Bandbreite nutzbar ist.
- Congestion Avoidance bedeutet, dass sich die Datenrate wieder um ein Vielfaches verringert, falls es zu einem Paketverlust kommt. Da es besonders bei WLAN oder Mobilfunknetzen des öfteren zu Paketverlusten kommen kann ist dieser Aspekt besonders hervorzuheben, denn er verzögert das Erreichen der maximal möglichen Datenrate.

Slow Start bedeutet also aus Sicht der Performance, dass bei einer neuen TCP Verbindung nicht die maximale Bandbreite zu Verfügung steht. Bei größeren Dateien ist zwar durch das exponentielle Wachstum das Maximum schnell erreicht, gerade aber bei kleineren mit wenigen Kilobyte ist dies oft nicht der Fall.

5.1.5 Above The Fold

Damit ist der auf einem Bildschirm sichtbare Bereich vor dem Scrollen gemeint. Diesem Bereich wird eine besondere Wichtigkeit zugesprochen.

„In an analysis of 57,453 eyetracking fixations, we found that there was a dramatic drop-off in user attention at the position of the page fold. Elements above the fold were seen more than elements below the fold: the 100 pixels just above the fold were viewed 102% more than the 100 pixels just below the fold.“ (Schade 2015)

Wichtige Informationen oder Navigationselemente sind meistens dort zu finden. Eine Webseite die nach dem Paradigma des Responsive-Webdesign aufgebaut ist kann dabei 3 oder mehrere Ansichten haben die alle einen unterschiedlichen „above the fold“ Bereich haben. Eine Anwendung



Abbildung 5: Darstellung des sichtbaren Bereichs vor dem Scrollen

kann aber auch unterschiedliche Seiten haben, auf dem der Anwender beim Aufrufen der Seite landen kann. Zum Beispiel wenn dieser An- oder Abgemeldet ist. Paradebeispiel dafür sind Facebook oder Twitter.

5.1.6 Kritischer Rendering-Pfad

Auf Englisch „critical render path“ genannt, ist der wohl wichtigste Begriff, wenn es um schnelle Ladezeiten geht. Durch die Optimierung des Rendering Pfads kann die benötigte Zeit für das erste Rendern der Seite erheblich verkürzt werden. Das Verständnis des Rendering-Pfads ist zudem eine wesentliche Voraussetzung für die Erstellung von schnellen Webanwendungen und soll in diesem Abschnitt ausführlich erklärt werden. Dabei wird er Begriff in seine zwei Teile Zerlegt: Kritischer und Rendering-Pfad.

Der Rendering-Pfad setzt sich aus den für die Anwendung nötige Ressourcen zusammen. Webanwendungen bestehen meist aus mehreren ausgelagerten Javascript und CSS Dateien die benötigt werden um überhaupt mit dem Rendern zu beginnen. Ein Seitenaufruf erzeugt folgenden Prozess:

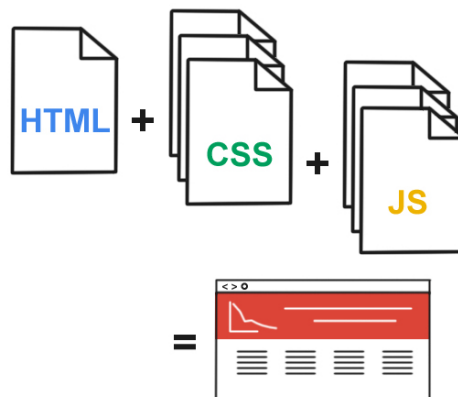


Abbildung 6: Ressourcen die für das Rendern von nöten sind

1. Nachdem die Adresse per DNS-Lookup (siehe Punkt: 5.1.2) in eine IP Adresse übersetzt wurde, wird eine TCP Verbindung hergestellt und der Browser beginnt damit die HTML

- Datei vom Server zu laden.
2. Der Browser liest das HTML von oben nach unten und sieht nach, ob zusätzliche Ressourcen benötigt werden (Bilder, Javascript oder CSS).
 3. Diese Ressourcen werden beim Server angefragt.
 4. Der Browser kann mit dem Rendern der Seite noch nicht beginnen, da er noch auf den Abschluss des Downloads wartet.
 5. Der Browser liest das Javascript und sieht nach, welche CSS Selektoren auf das HTML Dokument passen.
 6. Der Browser stellt fest, dass die Seite nun Angezeigt werden kann und beginnt mit dem Rendern.

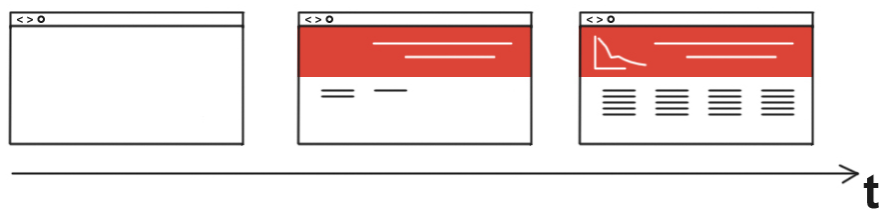


Abbildung 7: Der Render Prozess der im Browser abläuft

Wie im Prozessablauf abgebildet, ist ein frühes Rendering oft nicht möglich, weil Javascript und CSS das Rendern dadurch Blockieren, da sie zuerst Herunterladen werden und dann Übersetzt werden müssen, bevor der Browser entscheiden kann, dass die Seite gerendert werden kann. Kritischer Rendering-Pfad bedeutet, dass dem Browser die Daten zuerst zur Verfügung gestellt werden, die nötig sind um den „above the fold“ Inhalt zu rendern. Alle anderen Ressourcen sollten erst nach dem Laden der Seite Heruntergeladen werden um das Rendern nicht zu verzögern.

5.2 Perceived Performance

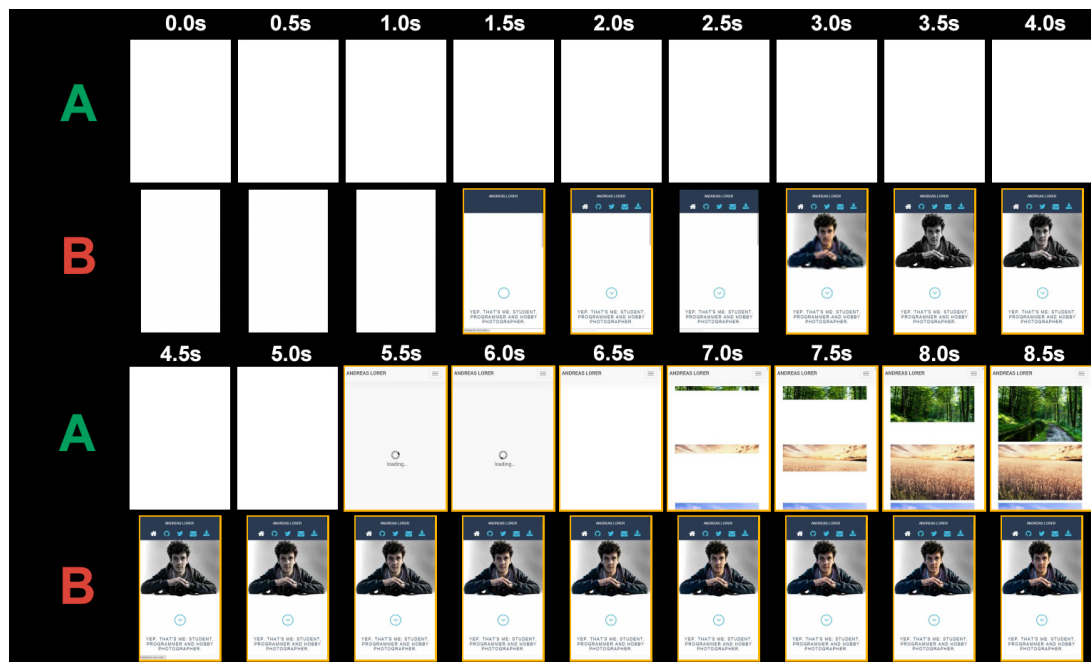


Abbildung 8: Zwei Seiten im Vergleich

Abbildung 8 zeigt die Seiten A und B, mit nahezu identischer Ladezeit. Der Unterschied besteht darin, dass Seite B bereits nach 1.5 Sekunden eine erste visuelle Rückmeldung für den Anwender zu sehen ist wohingegen Seite A erst nach 5.5 Sekunden dem Anwender zeigt, dass sie überhaupt ladet. „Perceived Performance“ steht also für die Zeit bis ein erste visuelle Rückmeldung für den Anwender zu sehen ist und bedeutet, dass die Ladezeit als schneller empfunden wird, als es eigentlich laut Messwerten der Fall ist. Warum diese „Perceived Performance“ für eine Webanwendung so wichtig ist zeigen mehrere Studien, deren Daten in folgender Infographik aufbereitet sind.

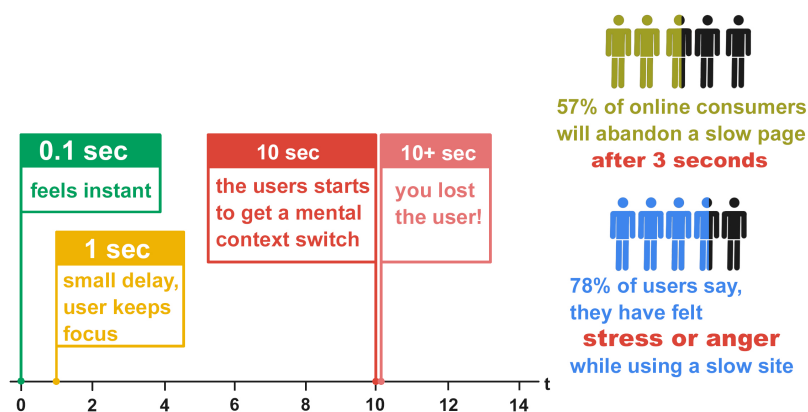


Abbildung 9: Einfluss und Effekt einer langsamen Seite auf den Anwender (Radware 2014, p. 8)(Nilsen 1993)

Bereits kleine verbesser- oder verschlechterungen der Ladezeit können einen großen Einfluss in auf den Anwender haben. Yahoo hat herausgefunden, dass wenn eine Seite um nur 400 Milisikunden schneller ist, sich der Traffic um 9% erhöhte.(Stefanov 2008) 57% der Online Konsumenten haben eine Seite, die länger als 3 Sekunden ladet bereits wieder verlassen. 78% der Anwender empfinden sogar Zorn oder Stress wenn eine Seite nicht Ladet oder es nicht ersichtlich ist, dass die Seite reagiert.

5.3 Latenz

TCP Handshake TCP Slowstart Http/1.1 Payload Einfluss der Datenrate auf die Geschwindigkeit

6 Entwicklung

6.1 Tools

6.1.1 Google Chrome Developer Tool

Stalled / Waiting DNS Lookup TTFB Start render Document Complete Document Loaded Audits Timeline

6.1.2 Webpagetest

Was ist Webpagetest Speedindex Features

6.1.3 Google Spreadsheet

6.2 Ausgangspunkt

6.3 Prozess der Suche

6.4 Prozess der Validierung

7 Best-Practices

7.1 Serverseitig

7.1.1 Verringern von DNS Lookups

7.1.2 HTTP Requests

pro / contra von sprites und großen concatenated files

7.1.3 Caching

pro / contra / problem bei updates

7.1.4 Pagespeed Mod

7.1.5 Images

5. Most sites fail to leverage best practices for optimizing images. Despite the fact that images represent one of the single greatest performance challenges (and opportunities), 34% of pages failed to properly implement image compression, and 76% failed to take advantage of progressive image rendering. (<http://www.radware.com/assets/0/314/6442478110/c810eee1-e86f-438a-b82f-3ad002bf1c75.pdf>)

7.2 Clientseitig

8 Workflow

8.1 Performanten Code schreiben

8.2 Minify und Uglify

8.3 Concatenating

8.4 Task Manager

8.5 Dependency Manager

8.6 Generators

Literatur

- [Goo10] Google. *Using site speed in web search ranking*. Website. 2010 (siehe S. 2).
- [Gri13] Ilya Grigorik. *High Performance Browser Networking*. <http://chimera.labs.oreilly.com/books/1230000000545/ch02.html> [Aufgerufem am 02.03.2015]. E-Book. 2013 (siehe S. 6).
- [Hol10] Urs Holzle. *Velocity 2010: Urs Holzle*. <http://tinyurl.com/px7m64m> [Aufgerufen am 27.02.2015]. Video from Velocity Conference. 2010 (siehe S. 2).
- [ItW15] ItWissen.info. *Shared Hosting*. <http://www.itwissen.info/definition/lexikon/Shared-Hosting-shared-hosting.html> [Aufgerufen am 26.02.2015]. 2015 (siehe S. 4).
- [Nil93] Jakob Nilsen. *Response Times: The 3 Important Limits*. Techn. Ber. <http://www.nngroup.com/articles/response-times-3-important-limits/> [Aufgerufen am 27.02.2015]. Nielsen Normand Group, 1993 (siehe S. 9).
- [Rad13] Radware. *Radware Mobile Infographic*. Website. http://blog.radware.com/wp-content/uploads/2013/11/Radware_SOTU_Fall_2013_Mobile_Infographic_Final1.jpg [Aufgerufen am 15.01.2015]. 2013 (siehe S. 2).
- [Rad14] Radware. *STATE OF THE UNION - Ecommerce Page Speed and Web Performance*. <http://www.radware.com/assets/0/314/6442478110/c810eee1-e86f-438a-b82f-3ad002bf1c75.pdf> [Aufgerufen am 27.02.2015]. Seite 8. 2014 (siehe S. 9).
- [Sch15] Amy Schade. *The Fold Manifesto: Why the Page Fold Still Matters*. Techn. Ber. <http://www.nngroup.com/articles/page-fold-manifesto/> [Aufgerufen am 26.02.2015]. Nielsen Normand Group, 2015 (siehe S. 6).
- [Ste06] Guido Stepken. *Anti-Pattern in der Softwareentwicklung*. <http://www.little-idiot.de/teambuilding/AntiPatternSoftwareentwicklung.pdf> [Aufgerufen am 26.02.2015]. 2006 (siehe S. 4).
- [Ste08] Stoyan Stefanov. *Exceptional Website Performance with YSlow 2.0*. <http://de.slideshare.net/stoyan/yslow-20-presentation> [Aufgerufen am 27.02.2015]. Slide Nummer 4. 2008 (siehe S. 10).
- [Ste11] Stoyan Stefanov. *Book of Speed*. <http://www.bookofspeed.com/chapter3.html> [Aufgerufen am 02.03.2015]. siehe Abbildung 3.4 - The-three-way handshake. 2011 (siehe S. 5).
- [t3n15] t3n. *Ist deine Website „mobile-friendly“? – Google steigert Druck auf Webmaster*. <http://t3n.de/news/google-mobile-friendly-589402/> [Aufgerufen am 25.02.2015]. 2015 (siehe S. 2).
- [TNS14] Google TNS Infratest BVDW. *Global Connected Consumer Study*. Website. <http://www.netzproduzenten.de/wp-content/uploads/2014/08/global-connected-consumer-studie-deutschland.pdf> [Aufgerufen am 14.12.2014]. 2014 (siehe S. 2, 3).
- [Web08] WebSiteOptimization.com. *The Psychology of Web Performance*. <http://www.websiteoptimization.com/speed/tweak/psychology-web-performance/> [Aufgerufen am 25.02.2015]. 2008 (siehe S. 2).
- [wik14] wiki.ubuntuusers. *Der Benutzer root*. <http://wiki.ubuntuusers.de/sudo> [Aufgerufen am 26.02.2015]. 2014 (siehe S. 4).

- [Wik15] Wikipedia. *Entwurfsmuster*. <http://de.wikipedia.org/wiki/Entwurfsmuster> [Aufgerufen am 26.02.2015]. 2015 (siehe S. 4).